

University of Alberta

PRACTICAL ISSUES IN NON-LINEAR SYSTEM IDENTIFICATION

by

Giti Esmaily Radvar



A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

in

Process Control

Department of Department of Chemical and Materials Engineering

**Edmonton, Alberta
Spring 2002**



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*

Our file *Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-69705-3

University of Alberta

Library Release Form

Name of Author: Giti Esmaily Radvar

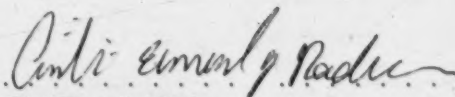
Title of Thesis: Practical Issues in Non-linear System Identification

Degree: Master of Science

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



Giti Esmaily Radvar

CME 536

University of Alberta

Edmonton, AB

Canada, T6G 2G6

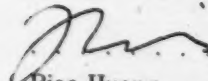
Date:

January 4, 2002

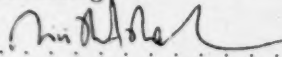
University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Practical Issues in Non-linear System Identification** submitted by Giti Esmaily Radvar in partial fulfillment of the requirements for the degree of **Master of Science** in *Process Control*.



Biao Huang



Sirish L. Shah



Qing Zhao

Date: *January 4, 2022*

To My Brother, Cyrus, in appreciation of his advise and support

Abstract

With the introduction of control to the process industries, it becomes a requirement to measure a process variable at a rate suitable for real-time control. However, due to high costs of on-line analyzers the measurement of the key variable is either unavailable or has long analysis cycle time. To cope with these trends, inferential modeling that extracts the relationships between known and unknown variables has been proposed. This includes methods based on visualizing the physical principles of the system (known as First Principles Modeling (FPM)) and data driven approach (such as partial least squares (PLS), group method of data handling (GMDH), artificial neural networks (ANNs) and model on demand (MOD)). Since most industrial plants exhibit a certain degree of non-linearity, usually a nonlinear model which captures the underlying behavior of the process, is required. Recent studies indicate that only an appropriate input/output data can validly identify the nonlinear process model.

In this thesis, the use of chemical engineering principles in characterization of the industrial chemical processes is explored. Utilizing the black box methodologies, these tools have been used for applications related to process modeling and soft sensor estimation when exact knowledge of the system is unavailable. In certain situations, these methods have provided a robust alternative to the complex physical modeling approach. Besides describing the theory, the applications of these techniques have also been investigated.

A number of issues in experimental design including designing an appropriate input sequence are studied. Extensive simulations and experiments are used to establish the superiority of multi level uniformly distributed input signal over other popular excitation signals.

The model on demand approach has been employed in identification of the pilot scale CSTR reactor. The domain of applicability of this approach has been extended to a multi rate system. The simulations and laboratory experiments are included wherever appropriate.

Acknowledgements

I would like to give my sincere thanks to my supervisor, Dr. Biao Huang. His experience and ingenuity was an invaluable asset throughout my master's program. His thoughts have enriched my thesis and with him, my confidence has been re-born. To me, he is not just an advisor, but a generous friend.

Dr. Sirish Shah deserves appreciation as one of the best teachers I have ever had. Dr. Fraser Forbes has my complete gratitude. His experience and knowledge played a very important role during my TA work.

I would like to offer my thanks for the invaluable support, guidance and motivation provided by my supervisors and friends at Syncrude Canada Ltd.: Edgar, Ahmed, Jack, Lalo, and the rest of upgrading unit.

It has been my privilege to work as a member of a great computer process control group at the University of Alberta. A sincere thanks to my student colleagues: Ashish, Ramesh, Arun, Harigopal, Shoukat, Zhengang, Sheng, Jianping, Xin, Raghu, Vikas, Folake, Monjour, Salim and others for providing a stimulating environment. Special mention must be made of Arun (his spiritual advise never be forgotten) and my office partner, Folake, for her encouragement. A word of thank also to the faculty and staff of the Department of Chemical and Materials Engineering for the help and resources.

I wish to express my heartfelt appreciation to my parents and rest of my family. Because of their spiritual support and encouragement, I have been able to finish my study at an advanced academic level. I also want to thank my wonderful husband because of his sublime love.

I would like to thank God for lifting my spirit when I am sad, being my companion when I am lonely, sharing my happiness and for being my most intimate friend. Thank Him for giving me good parents, husband, family and a great supervisor. It would not have been possible to complete this master's program without his grace.

Contents

1 Introduction	1
1.1 Scope and organization of this thesis	2
2 Inferential Modeling of a Multivariable Process using First Principles of Chemical Engineering	4
2.1 Introduction	5
2.2 Development of an industrial inferential model	6
2.2.1 Application of First Principles Technique to the Estimation of Fluid Coker Top Distillation Point	6
2.3 Process Description	6
2.3.1 Fluid Coker	6
2.3.2 Reaction section process	7
2.3.3 Scrubber section Process	8
2.4 Formulating Semi Physical Principles at Scrubber Overhead	8
2.4.1 Flow calculation of the products	8
2.4.2 Calculating the molecular weight of the products	11
2.4.3 Defining the equilibrium pressure and temperature at the top of the scrubber	11
2.4.4 Estimating the 95% cut point for combined gas oil (CGO)	13
2.5 Prediction results of the applied SFPM in estimating the 95% cut point CGO	15
2.6 Conclusion	16
3 Comparative Study of Empirical Multivariate Modeling Techniques and Industrial Application	17
3.1 Abstract	17
3.2 Introduction	18
3.3 Algorithm description of PLS	19

3.4	Industrial case studies	21
3.4.1	Application of PLS to the estimation of top distillation point in the Fluid Coker	21
3.4.2	PLS model for product flow gas at the scrubber overhead	23
3.4.3	Dynamic version of PLS model for product flow gas at the scrubber overhead	24
3.4.4	Comparison of PLS and PFPM models in estimating the scrubber top distillation point	24
3.5	Introduction to the GMDH technique	25
3.6	GMDH algorithm	26
3.7	Application of GMDH in estimating the overhead scrubber distillation point	28
3.7.1	GMDH analysis	29
3.7.2	Comparison of GMDH and PLS approach in predicting scrubber overhead distillation point	30
3.8	Introduction to Neural Networks	30
3.9	Fundamentals of Neural Networks and Back Propagation Algorithm	33
3.9.1	Simple Neurons and Networks	33
3.9.2	Back Propagation Algorithm	34
3.9.3	Neural Network analysis	35
3.9.4	Comparison of prediction result from GMDH and Neural networks algorithms	35
3.10	Conclusion	36
4	Issues in Experimental Design for Nonlinear System Identification	43
4.1	Abstract	43
4.2	Introduction	43
4.3	Pilot Scale Continuous Stirred Tank Reactor	44
4.4	Designing the Input Signal	46
4.5	Illustrative Example In Issues of Excitation of the Input Sequence	47
4.5.1	Experiment I: Applying Sum of Sine Waves as an Input	48
4.5.2	Simulation I: Applying Multi-Level Input Signal	49
4.5.3	Experiment II: Applying Multi-level Input Sequence to the Real Process	53
4.6	Estimation of Conductivity in the Continuous Stirred Tank Reactor, Performing Genetic Algorithm	54

4.7 Conclusion	57
5 Data Driven Model-on-Demand Approach	59
5.1 Abstract	59
5.2 Introduction	60
5.3 Theory	60
5.3.1 Selecting the Weights and Distance Function	61
5.3.2 Bandwidth	62
5.3.3 Classical Methods, Variance/Bias Balance	62
5.3.4 Model-on-demand Algorithm	63
5.4 Illustrative Examples of MOD Algorithm	63
5.4.1 Identifying the simulated CSTR system using a linear model	63
5.4.2 Evaluating MOD robustness using contrived nonlinear model structure	64
5.4.3 MOD prediction performance in the presence of disturbances	64
5.4.4 Comparison of prediction performance of CSTR system using ANNs, GMDH and MOD technique	65
5.4.5 Evaluation of MOD technique using experimental data	65
5.5 Multi-rate system	66
5.6 Simulated examples using multi rate operations	68
5.7 Conclusion	69
6 Conclusions	73
6.1 Contribution of this thesis	74
6.2 Recommendation for future work	74
A Nomenclature List	79
B Introduction to Genetic Algorithms	80
B.1 Performing Symbolic Regression, an Application of Genetic Algorithm	80

List of Figures

2.1	Schematic diagram of Fluid Coker	9
2.2	Prediction of 95% distillation point combined gas oil. The solid lines are the output from lab and the dots represent the model predictions.	15
3.1	Weightings of the first and second PC in variable selection	23
3.2	Predictions obtained using static PLS based empirical model on the training data set. The solid lines are the output from lab and the dots represent the model predictions.	24
3.3	Predictions obtained using static PLS based empirical model on the cross validation data set. The solid lines are the output from lab and the dots represent the model predictions.	25
3.4	Predictions obtained from static PLS based empirical model on the training data set. The solid lines are the actual output, product flow gas (FC921), and the dots represent the model predictions. FC921A represents the existing regression based model.	27
3.5	Predictions obtained from static PLS based empirical model on the cross validation data set. The solid lines are the actual output, product flow gas, and the dots represent the model predictions. FC921A represents the existing regression based model.	28
3.6	Predictions obtained using dynamic PLS based empirical model on cross validation data set of product flow gas. Both X and Y block are time shifted. The solid lines are the actual output and the dots represent the model predictions.	30
3.7	Comparison of PLS and PFFM predictions of scrubber overhead distillation point. The solid lines are the actual output. The dots and dashes represent the PFFM and PLS estimations respectively.	31
3.8	Basic scheme of propagations of variables in GMDH approach	32
3.9	Generalization of GMDH algorithm	33

3.10	Input to the GMDH algorithm (one output Y1 and m independent variables).	34
3.11	Construction of the new array Z.	35
3.12	Stopping criterion.	36
3.13	Prediction of 95% cut point CGO using group method of data handling empirical model on the testing data set. The solid lines are the output from lab and the dots represent the model predictions.	37
3.14	Comparison of GMDH and PLS prediction result in estimating the 95% cut point CGO. The solid lines and dots are the actual output from lab and the model predictions from GMDH respectively. The dashes represent PLS estimation.	38
3.15	Feedforward neural network with one hidden layer.	38
3.16	Flow chart of back propagation algorithm in three layer feedforward artificial neural networks.	39
3.17	Prediction of 95% cut point CGO using feedforward neural networks with back propagation learning algorithm on the training set of data. The solid lines are the actual output and the dots represent the model predictions. . .	40
3.18	Prediction of 95% cut point CGO using feedforward neural networks with back propagation learning algorithm on the testing set of data. The solid lines are the actual output and the dots represent the model predictions. . .	41
3.19	Prediction of 95% cut point CGO using NNs and GMDH. The solid lines are the actual output, the dots represent the model predictions from GMDH and the dashes are the estimation from NNs.	42
4.1	Schematic diagram of CSTR reactor	45
4.2	Illustration of frequency calculation using a Bode diagram	48
4.3	Modified Uniformly Distributed M-level Sequence	50
4.4	Process data for the SISO identification. The uniformly distributed multi-level input signal changes its value between maximum and minimum rate, 60-80 mL/min	50
4.5	Prediction results obtained from Linear regression and GMDH tool using an 8-level uniformly distributed input signal. The solid lines are the simulated output from the CSTR and the dots represent the model prediction.	52

4.6	Prediction results obtained from Linear regression using the sampling interval equal to 200 s. The solid lines are the output from the simulation and the dash line represents the model prediction.	54
4.7	Plant Data for identification of CSTR process.	55
4.8	Comparison of simulation and real experimental data. Solid line presents the actual data and the dash lines show the data obtained from simulation. . .	55
4.7	Prediction results obtained from experimental data. The solid line is the normalized actual output and the dots represent the estimated values from equation 4.9.	57
4.8	Prediction results obtained from experimental data. The solid line is the normalized actual output and the dash lines represent the estimated values from equation 4.9	58
4.9	Prediction results obtained from the non-linear model provided by the Genetic Algorithm. The solid lines are the output from the CSTR and the dash lines represent the model prediction.	58
5.1	Prediction results of NAOH conductivity obtained from global linear regression and MOD tool . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.	64
5.2	Prediction results of NAOH conductivity obtained from global regression and MOD tool using a false model structure . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.	65
5.3	Prediction results of NAOH conductivity from global linear regression and MOD tool applied on the noisy data set . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.	66
5.4	Prediction results of NAOH conductivity from global regression and MOD tool using a false model structure applied on the noisy set of data. The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.	67

5.5	Prediction results of NAOH conductivity from GMDH, MOD and ANNs tool applied on the noisy set of data. The solid lines are the simulated output from the CSTR; the GMDH and MOD are shown with dashes and the dots respectively. The square sign represents ANNs.	69
5.6	Prediction results of NAOH conductivity from MOD tool applied on the experimental set of data. The solid lines are the output from the pilot scale CSTR reactor. The solid lines are the actual output from the CSTR and the dots represent the MOD based prediction.	70
5.7	Prediction results of NAOH conductivity obtained from global linear regression and MOD tool applied on the multi rate system . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.	71
5.8	Prediction results of NAOH conductivity from global linear regression and MOD tool applied on the noisy data set. The data is obtained from the simulated multi rate system . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.	72
B.1	Representation of numeric expressions using tree structures. (Parent Population)	81
B.2	Example of crossover operation.	82
B.3	Example of Mutation operation.	82
B.4	Basic algorithm of the evolutionary computation, genetic algorithm in a flow chart scheme.	83

List of Tables

3.1	Process variables for the Fluid Coker	22
3.2	Process variables for the product flow gas	26
3.3	Commutative Y variance of static and dynamic PLS models.	29
3.4	Comparison of the black box techniques with first principles model	37
4.1	Summary of identification results for the CSTR reactor	49
4.2	Summary of simulation results for the CSTR reactor, using an eight level uniformly distributed signal	51
4.3	Summary of simulation results for the CSTR reactor, changing the sampling time to 50 s.	52
4.4	Summary of simulation results for the CSTR reactor, changing the sampling time to 100 s.	53
4.5	Nominal conditions for process variables.	56
4.6	Summary of identification results for the CSTR reactor, using an eight level uniformly distributed signal.	56
5.1	Comparison of mean square error (MSE) of different algorithms applied to data set with band limited white noise	68
5.2	Comparison of mean square error (MSE) of the discussed scenarios	70

Chapter 1

Introduction

Extracting the relationship between measured and unmeasured process variables (usually known as soft sensor development) has been the subject of system identification and process control for many years. One important technique for building these relationships is the use of physical principles including energy, momentum and mass balances. This tool referred as first principles modeling (FPM), has been applied successfully in different chemical systems. However, the first principles approach is based on a foundation of exact information and thorough understanding of the process which frequently requires a high amount of time and exploratory study and comes at a high cost. Although a failure to providing the complete process information does not invalidate this approach, usually lack or missing information provides an uncertainty and inaccuracy in the developed model. This problem is particularly significant when we deal with complex nonlinear systems.

To deal with the problem of inferring relationships between input and output data when very little *a priori* knowledge is available, the use of black box tools is proposed. There is a rich and well-established theory for black-box modeling of linear systems (see e.g. Ljung (1987) [34]; Söderström and Stoica (1989) [48]). It was not until the last few years that modeling and identification of nonlinear systems attracted wide interest in the control community. However, nonlinear modeling has been studied for a long time in the Statistics community, where it is known under the label *non-parametric regression*. This area is quite rich and numerous methods exist [27]. From an application point of view, these methods provide a good prediction when an appropriate model structure which best describes the behavior of the system, is available. One possible solution in choosing the structure of the model is the application of black box tools in combination with first principles technique. This usually outperforms the conventional models. Partial least squares or PLS regression is one of the applied tools which has been found to be more valuable in numerous situations

when the algorithm is amalgamated with the physical process information. There are various search techniques, on the other hand, which organize the model structure automatically when the mechanistic equation or any other first principles information of the process is completely unavailable. Typical examples of those tools are group method of data handling (GMDH) and Genetic Algorithms (GAs). Artificial neural networks (ANNs) offers another set of nonlinear identification tools which have attracted a lot of interest in the last two decades and is independent of the knowledge of the process. However, the primary complaint about this approach is that there is no explicit model structure available.

All these algorithms known as global techniques compress all available information into a compact model. Therefore, these methods become less attractive to deal with, when the number of data points increases and the associated optimization problem becomes more complex. Another alternative tool in nonlinear system identification, which overcomes this problem, is a nonparametric estimation called as model on demand (MOD). In MOD estimation all observations are stored on a database and the models are built "on demand" as the actual need arises [8]. This technique uses only relevant and a small portion of data to predict the key variable and determine a model as needed.

When dealing with nonlinear system identification, a number of issues concerning the experimental design and type of excitation signal should be considered. In past years, significant advances have been achieved in the field of experimental design for input signals, which deals with the identification of a linear system. However, there are still many unsolved problems with respect to nonlinear processes. There are various types of input sequences which have been applied in the development of a nonlinear model such as random binary signal 'RBS', M-level sequences and constant switching pace-signal [6]. However, the exposition of the type of input signals in nonlinear system identification is not exhaustive and requires further study.

1.1 Scope and organization of this thesis

This work is intended to serve as a study of nonlinear system identification in the field of chemical engineering. It involves the following specific objectives presented in sequential chapters:

- 1. Application of first principles approach to chemical engineering processes:
Here specific attention is given to the industrial case study at Syncrude Canada Ltd., in which some physical principles such as mass balance and flash calculation have

been employed.

- 2. Application of black box techniques in soft sensor development:

Employing some statistical and mathematical tools in exploring the relationship of the time series data is a common alternative when *a priori* knowledge of the system is unavailable. Focusing on the same industrial application, the feasibility of the various applied tools namely partial least square (PLS) regression, group method of data handling (GMDH) and neural networks (NNs) are investigated.

- 3. Issues of experimental design in identification of a nonlinear system:

As explained before, the problem of design of excitation signal has to be tackled before a robust model can be arrived upon. There exists some open problem in the literature on this issue and this requires further study. In this work, the formulation of the excitation signal proposed in [44] is modified and a new input sequence is proposed. This input is able to capture the underlying behavior of the nonlinear CSTR reactor.

- 4. Model on demand approach and its application to a multi rate system:

The initial thrust of this work is to review the nonparametric estimation of the nonlinear CSTR system on basis of observed data and model on demand (MOD) approach. Compared to global methods the advantage of MOD technique is that the modeling is optimized locally and therefore the prediction performance increases significantly. The domain of applicability of this approach has been extended to a multi rate system. The simulation results illustrate the effectiveness and feasibility of the algorithm.

Chapter 2

Inferential Modeling of a Multivariable Process using First Principles of Chemical Engineering

Abstract

In this chapter the issue of the first-principles based approach in system modeling, namely stating first principles balance equations of a chemical process, is discussed. The theoretical aspects and required knowledge of building a physical model are illustrated through an industrial case study. The ultimate goal is to estimate and develop an inferential model for the Fluid Coker top distillation point and enhance the performance of existing advance controller at Syncrude Canada Ltd. through this application.

However, due to a number of uncertainties and assumptions in the derivation, we do not intend to develop a very "accurate" first-principles based model. Instead, we try to exploit all available process information to search for an approximate model structure. The model parameters from this structure are then optimized to obtain a hybrid inferential model.

2.1 Introduction

In general, the measurement of key process variables at a rate suitable for real-time control is a problem in many industrial processes. Due to the high installation and maintenance cost of equipment, measurements are often unavailable. Even when appropriate instrumentation is available, on-line measurement is often restricted by long analysis cycle times. This often leads to the monitoring of key process variables through the use of off-line laboratory analysis and result in the process deviating from desired operating conditions with long disturbance recovery times, leading to unsatisfactory variability and a reduction in profit. Efforts towards alleviating these problems have included the development of inferential estimators. There are many variables measured on-line at relatively fast sample rates. These variables are indirectly related to the key 'difficult to measure' variables. Consequently, when an accurate inferential model is used on-line as a soft-sensor, the control performance can be improved due to the higher sampling rate.

One preferred methodology that may be adopted when building an inferential model is the 'bottom up' approach, which is based on formulating and solving material and energy balance on chemical process systems. This approach, usually referred to as the first principles modeling technique, introduces a fundamental engineering approach to solving process-related problems such as establishing the relations between known and unknown process variables. Despite various advantages of this technique in system modeling, stating the balance equations for mass, energy and momentum might not always lead us to an accurate and reliable process model. The primary complaint against first principles modeling (FPM) often cited is that applying physical principles in industrial application is not straightforward and requires a great deal of exploratory study and exact knowledge of the process. This is seldom available.

Although, a failure to provide the data measurements or any other required information does not invalidate this approach, the lack of information and measurements lead us to apply a large number of assumptions and employ various empirical equations and optimization factors which cause inaccuracy in the developed model.

To examine the discussed issues in inferential modeling using FPM, in the following section we provide one industrial case study at Syncrude Canada Ltd. to illustrate our point. This work uses the basic understanding of the process and follows the mass balance philosophy, Flash calculation and ASTM standard presented in M.Felder *et al.*, (1989), K. Sattler *et al.*, (1995) and the API handbook (1996) [1] respectively. The performance of

FPM inferential model is evaluated by making the comparison between the estimated value and actual off-line laboratory data from Syncrude Canada Ltd.

2.2 Development of an industrial inferential model

2.2.1 Application of First Principles Technique to the Estimation of Fluid Coker Top Distillation Point

The FPM technique has been employed to obtain an estimator for the top distillation point of Combined Gas Oil (CGO) in a Fluid Coker whose sampling interval associated with offline laboratory procedures is 12 hours. In this application, the Coker top temperature at which 95% of combined gas oil is distilled, is considered as the key quality variable.

Since the unique topic in all FPM problems is that, given values of some input and output stream variables calculate values of others, it is necessary to have a complete knowledge of the process to provide an accurate model. However, due to the complexity of the system and interrelation of Fluid Coker with other units, we are unable to have a fundamental and complete understanding of the process. Thus, to construct the physical model for the Fluid Coker, first we present an illustration of the Coker and other related systems in the upgrading unit. Using this information, an optimal model for the output variable, 95% cut point CGO has been developed. The unknown parameters are treated as optimization factors which are optimized subsequently. Due to reasons that the present model does not exactly follow the physical law of the process, in the rest of the study we use the name semi/pseudo first principles technique (SFPM) instead of (FPM).

2.3 Process Description

2.3.1 Fluid Coker

Upgrading of Bitumen to a low sulphur sweet crude oil is the main goal in the upgrading unit. This can be achieved by thermal cracking the Bitumen in a fluidized coke bed reactor at a high temperature. Under high temperatures, Bitumen is cracked into lighter hydrocarbon products and fluid coke. The products from the Fluid Coker are typically:

- Sour Fuel Gas,
- Untreated Naphtha,
- Combined Gas Oil,
- Burner Off-Gas,

- Product Coke, and
- Sour Water

The Fluid Coker consists of three different sections: reactor, burner and scrubber. In the reactor, Bitumen is dispersed into a fluidized bed of hot coke particles for cracking. The hydrocarbon vapors produced flow upward through the reactor into the scrubber. In the scrubber, the reaction products are cooled and the liquid product is recycled from the scrubber to the reactor for further cracking. The vapors from the scrubber pass through a Fractionator which produces the following products:

- Sour Fuel Gas
- Untreated Naphtha
- Untreated Combined Gas oil, CGO
- Light Gas Oil, LGO
- Heavy Gas Oil, HGO

2.3.2 Reaction section process

Fluid coking mechanism in the Fluid Coker shows that three reactions occur in the reaction section:

- The heavy hydrocarbon feed cracks into lighter products,
- Coke forms, and
- The hydrocarbon vapors and coke begin to separate.

The "cracking" that takes place in the reactor refers to a combination of reactions:

1. Cracking and condensation; removing side chains and breaking rings to make lower molecular weight, lower boiling range, higher hydrogen/carbon (H/C) ratio hydrocarbons (cracking).
2. polymerization; combination of radical groups giving an increase in molecular weight and boiling range.
3. Isomerization; Rearrangement within the molecule, no change in molecular weight or H/C ratio, little or no change in boiling range.

The reaction occurs in a fluid bed of coke under high temperature and is controlled by the flow of hot coke from the burner. The hydrocarbons, at this stage, either crack and vaporize or carbonize on the coke particle. The lighter, vaporized and cracked hydrocarbon travels upward through the dense bed to the dilute phase section of the reactor and the heavier coke particles move downward toward the stripping section.

2.3.3 Scrubber section Process

The scrubber is an extension of the reaction section and can be divided into three sections: pool, shed and grid sections. The scrubber has different functions dealing with:

- Cooling the hydrocarbon effluent of the reactor and separating it from entrained coke.
- Separating the hydrocarbon stream according to the boiling point with the lighter fractions going on into the fractionator while the heavier components are recycled into the reaction section.

A schematic diagram of the process is shown in figure 2.1.

2.4 Formulating Semi Physical Principles at Scrubber Overhead

To present a description of the approach, we break the problem into the following steps:

- Flow Calculation of the Products
- Calculating the Molecular Weight of the products
- Defining the Equilibrium Pressure and Temperature at the top of the Scrubber
- Estimating the key variable, 95% cut-point Combined Gas Oil (CGO)

In the following sections the description of each step is presented in more detail. The variables that are applied for building the SFP based inferential model are listed in each subsection. As the first assumption, it is assumed that the system is in steady state condition and all dynamics are negligible.

2.4.1 Flow calculation of the products

As already indicated, since the flow of the products from the scrubber is not directly measured the mass balance relationship that accounts for changes in the flow of the specified product is necessary.

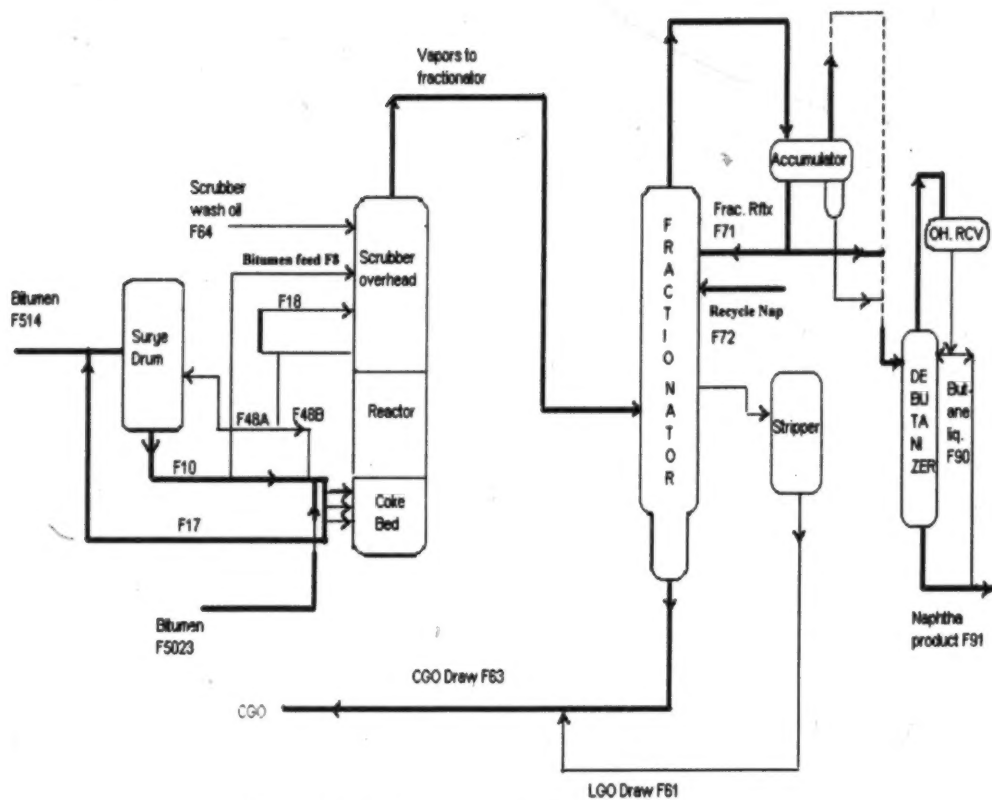


Figure 2.1: Schematic diagram of Fluid Coker

We follow the balance equation around the scrubber, fractionator and other related units and calculate the flow of Naphtha, Light Gas Oil (LGO), Heavy Gas Oil (HGO) and Steam.

Calculating the flow of product Naphtha

The estimation of flow of Naphtha involves the effluent and rerun Naphtha from/to the fractionator as follows:

Total flow of Naphtha from the Scrubber =

Effluent Naphtha from the fractionator - Rerun Naphtha to the fractionator

$$Naphtha_{Scrub} = F90 + F91 - (F72) * convf1 - F547 \quad (2.1)$$

Where,

F90 : Butane liquid product, KBPD

F91 : Naphtha product from frac., KBPD

F72 : Rerun Naphtha to the frac., BPD

F547: Frac. overhead plant 15, KBPD

convf1: conversion factor (0.001)

Naphtha_{Scrub}: KBPD

Calculating the flow of LGO and HGO

Balance equation for LGO and HGO can be formulated as follows:

Total flow of HGO/LGO from the scrubber =

Effluent flow of LGO/HGO from the fractionator

$$HGO_{Scrub} = F63 - F61 + F64 \quad (2.2)$$

$$LGO_{Scrub} = F61 \quad (2.3)$$

where,

F63: CGO product from the frac., KBPD

F61: LGO product from the frac., KBPD

F64: Scrubber wash oil (HGO) flow from the frac., KBPD

HGO_{Scrub}: KBPD

LGO_{Scrub}: KBPD

Calculating the flow of steam

Following balance equation shows the estimation of flow of steam from the scrubber overhead:

Total flow of steam from the scrubber=

Effluent sour water from the fractionator - inlet steam flow to the fractionator

$$Steam_{Scrub} = F132 * convf2 - (Fc70 + Fc65) * convf3 \quad (2.4)$$

where,

F132: Coker sour water to the tankage, GPM

Fc65: Steam to HGO stripper, KPPD

Fc70: Steam to LGO stripper, KPPD

convf2: Conversion factor (8.3454)

convf3: Conversion factor (1000/60)

Steam_{Scrub} : $\frac{LB}{MTN}$

Calculating the flow of gas at the scrubber overhead

The flow of gas at the top of the scrubber is calculated based on the Partial Least Square regression as described in Chapter 3. We use the available historical data for this variable and treat the scrubber top flow gas as a known parameter in the rest of this study¹.

2.4.2 Calculating the molecular weight of the products

In order to calculate the molecular weight of Naphtha or CGO, we require the knowledge of some physical properties such as specific gravity of hydrocarbons (API) and mean average boiling point (MABP).

To estimate the MABP, we follow the API handbook (1996) [1] and convert the simulated ASTM 2887 to ASTM D86 distillation to calculate the volumetric average boiling point (VABP). The VABP is then converted to MABP.

Having the MABP stated above, the molecular weight can be calculated as follows:

$$S = \frac{141.5}{131.5 + API} \quad (2.5)$$

$$M = 20.486 * [\exp(1.165 * 10^{-4} * T_{MeR} - 7.78712 * S + 1.1582 * 10^{-3} * T_{MeR} * S)] \dots \\ * T_{MeR}^{1.26007} * S^{4.98308} \quad (2.6)$$

where, S is the specific gravity, 60 F/60 F calculated from API and T_{MeR} is MABP based on degrees Rankine.

$$K = \frac{T_{MeR}^{\frac{1}{3}}}{S} \quad (2.7)$$

K: Watson characterization factor

2.4.3 Defining the equilibrium pressure and temperature at the top of the scrubber

As mentioned before, liquid and vapor mixtures in the scrubber are in constant contact and they reach the equilibrium point somewhere at the scrubber overhead. Therefore, the scrubber acts as a flash operation process where partial separation of the product from the reactor occurs. To estimate the vapor pressure of combined gas oil, we require the total pressure of the mixture as well as equilibrium temperature at the scrubber overhead. Since

¹See chapter 3 for more information on scrubber top flow gas calculation.

these variables are not directly measured, we make a number of assumptions to provide a tractable solution.

To estimate the vapor pressure of combined gas oil, the mole ratio at the scrubber overhead is required. To estimate the mole ratio of active hydrocarbons in the system, the flow of each product calculated before, is converted to *Lb/min* and divided to the molecular weight. The mole based total flow can be formulated as follows:

$$Mole_{total} = Naphthamole + Hgomole + Lgomole + Fuelgasmole + Steammole \quad (2.8)$$

Due to the simplification and assumptions in flow estimation, we add an optimization parameter in calculating the mole fraction to reduce inaccuracy and achieve better results in final model.

$$CGOmolefrac = (Lgomole + Hgomole) / (Mole_{total} * x(1)) \quad (2.9)$$

Note that $x(1)$ and $CGOmolefrac$ are optimization parameter and mole ratio respectively. Applying the flash calculation, we estimate the vapor pressure of CGO at the top of scrubber as follows:

$$\pi = P19 - P18 \quad (2.10)$$

$$P_{vapor-pressure} = \pi * CGOmolefrac * x(2) \quad (2.11)$$

$$T_{equilibrium}^2 = T_{508} * x(3) \quad (2.12)$$

where,

π = Total (vapor) pressure of the system

P19: Reactor dilute phase pressure

P18: Reactor scrubber differential pressure

T508: Temperature at the bottom of the shed section

$x(2)$: Mole fraction of CGO in the liquid phase which is an unknown parameter, treated as second optimization variable with the constraint $0 < x(2) < 1$

$P_{vapor-pressure}$: Calculated vapor pressure of CGO in PSI

$x(3)$: Optimization parameter

$$p^* = ((P_{vapor-pressure} + 14.696) * \frac{760}{14.696}) * x(4) \quad (2.13)$$

p^* : Vapor pressure of CGO in mmHg

$x(4)$: Optimization variable

2.4.4 Estimating the 95% cut point for combined gas oil (CGO)

Following the API handbook [1], the normal boiling point of pure hydrocarbons and narrow boiling petroleum fractions can be estimated by the following procedures:

$$\Delta T = T_b - T'_b = 2.5f(K - 12)\log \frac{P^*}{760}$$

$$X = \frac{\frac{T'_b}{T_{\text{eqe}}} - 0.002867(T'_b)}{748.1 - 0.2145(T'_b)}$$

$$\log p^* = \frac{3000.538X - 6.76156}{43X - 0.987672} \text{ for } X > 0.0022 \text{ (} P^* < 2\text{mmHG)} \text{}$$

$$\log p^* = \frac{2663.129X - 5.994296}{95.76X - 0.972546} \text{ for } 0.0013 \leq X \leq 0.0022 \text{ (} 2\text{mmHG} \leq P^* \leq 760\text{mmHG)} \text{}$$

$$\log p^* = \frac{2770.085X - 6.412631}{36X - 0.989679} \text{ for } X < 0.0013 \text{ (} P^* > 760\text{mmHG)} \text{}$$

Combining all these equations with respect to T'_b/T_b (unknown dependent variables in our case), we have:

$$T'_b = \frac{740.378 \log p^* - 4796}{\frac{36}{T_{\text{equilibrium}}} \log p^* - \frac{2770.08}{T_{\text{equilibrium}}} + 0.20188 \log p^* - 0.581326} \quad (2.14)$$

for $P^* > 760\text{mmHG}$

$$T'_b = \frac{727.56 \log p^* - 4484.33}{\frac{95.76}{T_{\text{equilibrium}}} \log p^* - \frac{2663.129}{T_{\text{equilibrium}}} + 0.18116 \log p^* - 0.5222} \quad (2.15)$$

for $2\text{mmHG} \leq P^* \leq 760\text{mmHG}$

$$T'_b = \frac{738.877 \log p^* - 5058.32}{\frac{43}{T_{\text{equilibrium}}} \log p^* - \frac{3000.538}{T_{\text{equilibrium}}} + 0.19952 \log p^* - 0.58975} \quad (2.16)$$

for $P^* < 2\text{mmHG}$

$$T_b = T'_b + 2.5f(k - 12) \frac{\log p^*}{760} \quad (2.17)$$

where,

$T_{\text{equilibrium}}$: Absolute equilibrium temperature from previous section

T'_b : Normal boiling point, in degrees Rankine

f : Correction factor. For all sub atmospheric vapor pressures and for all substances having normal boiling points greater than 400 F, $f=1$. For substances having normal boiling points less than 200 F, $f=0$. For super atmospheric vapor pressures of substances having normal boiling points between 200 F and 400 F, f is given by:

$$f = \frac{T_b - 659.7}{200}$$

k : Watson characterization factor.

T_b : True boiling point in degrees Rankine

$$TBP(100) = T_b - 459.67 \quad (2.18)$$

$TBP(100)$: True boiling point in degrees Fahrenheit

The correction of the assumption that the calculated T_b is equal to the true boiling point $TBP(100)$, is embedded in optimization factor $x(5)^3$. To compare the estimated temperature with the one obtained from the lab, we use the API handbook and convert the TBP to lab standard ASTM 2887 and estimate the simulated distillation $SD(100)$.

As stated in the API handbook, it can be assumed that the true boiling point temperature at 50 volume distilled is equal to the simulated distillation temperature at 50 weight percent distilled.

To have the relationship between the true boiling point and the cumulative percent volume of the evaporated product at each temperature, we assume that the distillation curve between 50 and 100 % is linear. To reduce the inaccuracy of this estimation, an optimization variable $x(5)$ is added to the calculated slope. It should be noted that simulated distillation 50 ($SD(50)$) is provided by the laboratory:

$$Slope = \frac{(SD(100) - SD(50))}{50} * x(5) \quad (2.19)$$

Making the linear interpolation between (50 and 100 %), the temperature at the required cut point can be estimated. To adjust the assumptions made in this calculation, the final correction factor $x(6)$ is added to the following equation.

$$T_{sd} = (Slope * (y - 50) + SD(50)) + x(6) \quad (2.20)$$

y : volume percent distilled (95% in this approach)

$$T_c = \frac{T_{sd} - 32}{1.8} \quad (2.21)$$

³see equation 19.

where, T_c is the calculated distillation point in $^{\circ}\text{C}$ at any cut point above 50%.

2.5 Prediction results of the applied SFPM in estimating the 95% cut point CGO

The proposed calculation for the inferential model has been implemented in MATLAB environment. Throughout this work, the nonlinear least squares optimization function has been utilized⁴. The unknown variables $x(1)$, $x(2)$, $x(3)$, $x(4)$, $x(5)$ and $x(6)$ have been optimized using 130 data points collected from April to June 2000.

A second set of 120 records collected from July to September 2000, was used for model validation. Figure 2.2 shows the actual and predicted 95% distillation point of combined gas oil.

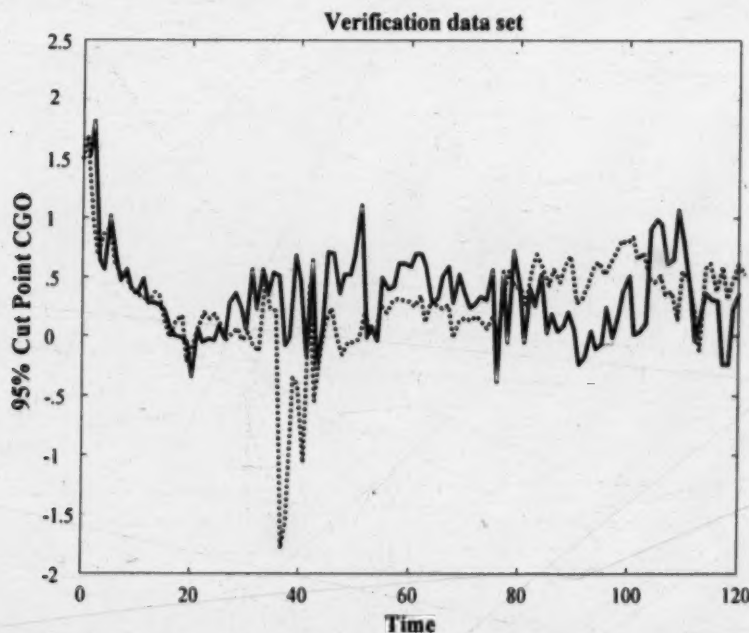


Figure 2.2: Prediction of 95% distillation point combined gas oil. The solid lines are the output from lab and the dots represent the model predictions.

⁴This function is defined as "lsqnonlin" in MAT LAB.

2.6 Conclusion

The technique of applying first principles laws on an industrial system, Fluid Coker has been explained in this chapter. This technique provides a semi-empirical inferential model to predict the objective variable, 95% cut point combined gas oil. The results reveal that this approach requires detailed knowledge about the physics and chemistry of the process. It is indicated that the development of a realistic model for such a complex system involves a number of simplifying assumptions, which may be subject to inaccuracies. Consequently, for such a practical application, it is important to include a number of optimization and correction factors. By optimizing all unknown variables of the model developed from the first principles, we obtained a reasonably good prediction of 95% distillation point of CGO. This technique provides an approximate model structure of the process, which might be useful for the development of an "optimal" and more "accurate" inferential model using other non-linear inferential approaches.

Chapter 3

Comparative Study of Empirical Multivariate Modeling Techniques and Industrial Application

3.1 Abstract

This chapter deals with the study of two nonlinear system identification methods, namely the Partial Least Square regression (PLS) and Artificial Feed-forward Neural Networks (ANNs). The group method of data handling (GMDH) is presented as another alternative technique to construct a nonlinear polynomial structure for a complex system when very little priori knowledge of the process is available.

Applying the PLS algorithm to the industrial case study, presented in the previous chapter, the inferential model is constructed to estimate the scrubber overhead distillation point. This technique is also extended to the dynamic estimation of the product flow gas from the Coker. We evaluate the feasibility and application of NNs and GMDH techniques by practically implementing these methods on the same case study, predicting the scrubber overhead distillation point. Comparison of the prediction results demonstrates the potential capability of all proposed algorithms in estimating the quality variable, 95% distillation point of combined gas oil.

3.2 Introduction

The problem we are addressing in this chapter is how to derive mathematical models and infer relationships between current and past input-output data and future outputs of a system when very little information about the process is available. This is known as black-box modeling [27].

In process simulation studies as well as process behavior prediction, good input-output representation of a process is of paramount importance. In addition, it is required that these models are interpretable, in the sense that, by analyzing the model there is an understanding of the process behavior. However, since most industrial plants are often complex, pure black box models might be unable to describe the process information and therefore have limited validity and may also contain uncertainty. This problem is particularly significant when one switches from linear to nonlinear system identification. Choosing the appropriate model structure which best describes the behavior of the system is the crucial part in nonlinear input-output modeling. To deal with this problem, it is proposed to use black box tools in combination with first principles technique. By using the first principles to determine the model structure and then estimating the unknown parameters from data, a hybrid black-box/principles model can be constructed. Such a model can outperform a "conventional" one. This remarks the fact that the combination of the physical and black box models, usually referred as "gray box technique", might lead us to a better prediction performance.

The purpose of this study is to give an exposition of presently available techniques of nonlinear modeling. It illustrates the basic principles as well as their application on the industrial data sets. As the first applied technique, we review the theory of PLS algorithm and employ this technique on the case study discussed in the previous chapter. We provide the steady-state model as well as dynamic empirical model for the scrubber overhead distillation point and the flow of product gas respectively. The structure of the former model is built upon the first principles of chemical engineering in a way that the variables and non linearities defined in the PLS model are those derived according to the first principles model. The accuracy of the prediction in this approach depends on how good is the model structure and how frequent observation points are collected. Thus having enough data points as well as a good model structure are the main requirements of this technique.

GMDH is another alternative black box technique discussed in this work. Utilizing the algorithm on the Coker data set, the structure of the model is searched automatically and a nonlinear polynomial model is produced which predicts the quality control variable in the Fluid Coker.

Providing the overview of mathematical theory to explain Feedforward Neural Networks, the application of this nonlinear identification tool is discussed next. This technique is applied on the same industrial data set. However the availability of large number of observations is the main requirement of this tool while there is no need of explicit model structure.

3.3 Algorithm description of PLS

In a multivariate system, when we deal with relatively fewer variables, which are not strongly correlated and have a well-understood relationship to the responses, multiple linear regression (MLR) can be a good way to turn the data into information. However, if we are faced with many variables and ill-understood relationships, MLR might be inefficient or inappropriate. In such a case, the approach of using empirical latent variable models is a good alternative to the classical multiple linear regression.

PLS is the generalization of principle component analysis (PCA) in a way that eliminates redundancies in the data blocks by extracting the latent factors, which account for most of the variation in the response. PLS has a conceptual similarity to canonical correlation analysis (CCA) in that the provided model explains the combinations of variables which are highly correlated [28]. From a practical point of view, PLS algorithm represents the dimensionality reduction and concentration of the variance of the process in the first few components.

In the PLS method, U block of independent variables is related to a Y block of dependent variables as follows:

$$Y = U\beta + E \quad (3.1)$$

where, E is the error or residual matrix.

The principal component transformations for the matrix U and Y are then given by:

$$T = UC \quad (3.2)$$

$$y = YD \quad (3.3)$$

The columns of T and y are called the scores vectors, and the columns of C and D are called weighting vectors subject to the constraints that $C'C = 1, D'D = 1$.

The first principle components of U and Y are represented by t_1 and y_1 respectively (" $t_1 = Uc_1, y_1 = Yd_1$ ").

As mentioned before, one goal in PLS is to maximize the covariance between U and Y block latent variables. Defining the Lagrangian multipliers, λ_1, λ_2 , we can express this property in an objective function as follows:

$$\begin{aligned} \max \text{covariance } (Uc_1, Yd_1) &= \max ((Uc_1)'Yd_1) \\ J_1 &= (Uc_1)'Yd_1 - \frac{1}{2}\lambda_1(c_1'c_1 - 1) - \frac{1}{2}\lambda_2(d_1'd_1 - 1) \end{aligned} \quad (3.4)$$

Differentiating equation 3.4 partially with respect to c_1', d_1' and equating the resulting equations to zero, we get:

$$\frac{\partial J_1}{\partial c_1'} = U'Yd_1 - \lambda_1 c_1 \quad (3.5)$$

$$\frac{\partial J_1}{\partial d_1} = Y'Uc_1 - \lambda_2 d_1 \quad (3.6)$$

$$U'Yd_1 = \lambda_1 c_1 \quad (3.7)$$

$$Y'Uc_1 = \lambda_2 d_1 \quad (3.8)$$

Solving equation 3.8 with respect to d_1 and substituting the solution into 3.7, we have:

$$d_1 = \frac{1}{\lambda_2} Y'Uc_1 \quad (3.9)$$

$$U'YY'Uc_1 = \lambda_1 \lambda_2 c_1 \quad (3.10)$$

From 3.10, we notice that c_1 is the corresponding eigenvector of $U'YY'U$.

To show that the product $\lambda_1 \lambda_2$ corresponds to the largest eigenvalues of $U'YY'U$, we look at the second derivative of the objective function:

From 3.5 and 3.9 we have:

$$\frac{\partial J_1}{\partial c_1} = \frac{1}{\lambda_2} U'YY'Uc_1 - \lambda_1 c_1 \quad (3.11)$$

To maximize the objective function, we make the second derivative of 3.11 with respect to c_1 . Note that the resulting equation should be smaller than zero.

$$\frac{\partial^2 J_1}{\partial^2 c_1} = \left(\frac{1}{\lambda_2} U'YY'U - \lambda_1 \right) < 0 \quad (3.12)$$

As a result: $U'YY'U < \lambda_1 \lambda_2$

In order to define the first component of Y score, y_1 , we regress Y based on the first score vector (Uc_1):

$$Y = (Uc_1)Z_1 + E_1 \quad (3.13)$$

where, $c_1 Z = \beta_1$.

Applying least square regression, we can estimate the coefficients $\hat{Z}_1$ and $\hat{\beta}_1$.

$$\hat{Z}_1 = ((Uc_1)'(Uc_1))^{-1} (Uc_1)'Y \quad (3.14)$$

Note, $t_1 = Uc_1$,

$$\hat{\beta}_1 = c_1(t_1't_1)^{-1}t_1'Y \quad (3.15)$$

Define, $\hat{Y}_1 = U\hat{\beta}_1$,

$$\hat{Y}_1 = Uc_1(t_1't_1)^{-1}t_1'Y = \frac{t_1 t_1' Y}{t_1' t_1} \quad (3.16)$$

Using the fact that $d_1' d_1 = 1$ we define:

$$d_1 = \frac{(\frac{t_1' Y}{t_1' t_1})'}{|\frac{t_1' Y}{t_1' t_1}|} \quad (3.17)$$

Then,

$$b_1 = |\frac{t_1' Y}{t_1' t_1}| \quad (3.18)$$

$$\hat{Y}_1 = t_1 b_1 d_1'$$

Applying the deflation process (model order reduction), we can compute the second elements of Y and X scores as follows:

$$\begin{aligned} Y_2 &= Y - \hat{Y}_1 \\ Y_2 &= (I - \frac{t_1 t_1'}{t_1' t_1}) Y \end{aligned} \quad (3.19)$$

At this stage, we choose t_2 in such way that $t_1' t_2 = 0$

$$\begin{aligned} t_2 &= U_2 c_2 \\ U_2 &= (I - \frac{t_1 t_1'}{t_1' t_1}) U_1 \end{aligned} \quad (3.20)$$

The procedure of determining other scores and loading vectors is continued until we achieve the required number of PLS dimensions. Final error matrices E_u and E_y will be:

$$E_u = U - (\frac{t_1 t_1'}{t_1' t_1} U_1 + \frac{t_2 t_2'}{t_2' t_2} U_2 + \dots + \frac{t_l t_l'}{t_l' t_l} U_l) \quad (3.21)$$

$$E_u = U - t_1 p_1' - t_2 p_2' - \dots - t_l p_l'$$

where $p_i = \frac{u_i' t_i}{t_i' t_i}$ is known as the loading factor.

$$E_y = Y - t_1 b_1 d_1' - t_2 b_2 d_2' - \dots - t_l b_l d_l' \quad (3.22)$$

For more information about PLS algorithm readers are referred to [17],[28],[24] and [31].

3.4 Industrial case studies

3.4.1 Application of PLS to the estimation of top distillation point in the Fluid Coker

The focus of this study is the Fluid Coker presented in the previous chapter. As discussed before, the large sampling intervals and time delays associated with offline laboratory procedures are the main hindrances to efficient control of the distillation point at the top of the scrubber.

It was shown that apart from the time consumed in developing the physical model, lack of information and measurements in the process made one apply various empirical equations and assumptions which result in uncertainty in the prediction.

The objective of this application is to build a data based inferential model for the 95% distillation point and to compare the results with the one developed in the previous chapter. In this application we use the Pseudo First Principles Model (PFFPM) to determine the important variables as well as the structure of the model and then apply partial least square technique to improve the performance of the model in predicting the controlled variable 95% distillation point.

Selection of important variables and structure of the models

In table 3.1 all variables along with their tag numbers that are applied in PFFPM are listed. Data on these variables are logged on to the database every 12 hours. In building the PLS based inferential model, the X block comprises of all listed variables while the Y block has one variable, the 95% cut point CGO. Due to the large sampling time, the dynamics are omitted. To select the important variables, a loadings plot involving the weights attached

Tag-name	Description	Tag-name	Description
F90	Butane liquid Production	F10	Fresh fee to the reactor
F91	Naphtha Production	F5023	VDU bottoms
F72	Rerun Naphtha to the Fract	F514	Coker feed+ Resid
F547	Fract overhead Plant 15	QX909	Scrubber OVHD heat
F63	CGO to tankage	P19	React. dilute phase press
F61	LGO flow rate	P18	React. scrubdiffrl Press
F64	Scrubber wash oil	T508	PA EX BTM SHED
F132	Cocker sour water	08100800.ML	API gravity of CGO
T13	Reactor temperature	T13*F5023	NON-LINEAR TERMS
F8	Bitumen feed to the scrubber	F5023*QX909	
F541	Top feed ring	P19-P18	
F542	Middle feed ring	LOG(P19-P18)	
F543	Bottom feed ring	LOG(P19-P18)*	
F18	Scrubber pool quench	T508	

Table 3.1: Process variables for the Fluid Coker

to the X variables is used. As it is shown in figure 3.1, the variables with small weightings on the predictions are considered as prime candidates for deletion.

At this stage, some flow rate and pressure variables have been dropped. The remaining 10

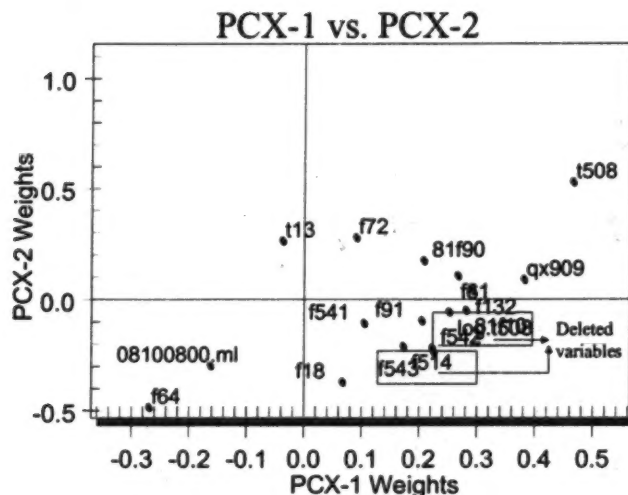


Figure 3.1: Weightings of the first and second PC in variable selection

variables are considered for further analysis. A final PLS model is constructed based on these important process variables. Three PLS components are sufficient to capture nearly most of the variation in response variable.

PLS analysis reveals an estimation of 95% cut point CGO, shown in figure 3.2 and 3.3.

3.4.2 PLS model for product flow gas at the scrubber overhead

It was stated in the last chapter that one of the key variables in calculating the mole fraction of active hydrocarbons used in the pseudo first principles model is the flow of effluent gas from the scrubber. Two different methods are currently used at upgrading unit of Syncrude Canada Ltd. to achieve the amount of this variable:

- The mass balance equation which gives reliable and accurate results but depends on a large number of other process measurements and the model structure is highly complex.
- The model obtained from ordinary regression analysis which has a very simple structure but gives very inaccurate results.

At this stage, to overcome this dilemma, we apply the PLS method along with the process knowledge to search for all important independent variables. To investigate which variables are highly correlated with the response variable, product flow gas, the weighting and variable importance plots were inspected. The selected variables are listed in table 3.2. After dropping the variables which have less contribution to the prediction of this objective variable, the PLS model is constructed for product flow gas. This model was presented in

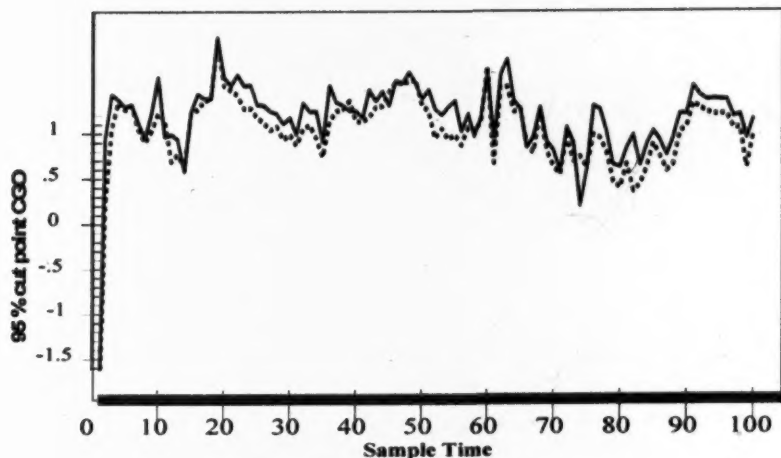


Figure 3.2: Predictions obtained using static PLS based empirical model on the training data set. The solid lines are the output from lab and the dots represent the model predictions.

chapter 2. Figure 3.5 shows the effectiveness of PLS approach compared to the normal regression method.

3.4.3 Dynamic version of PLS model for product flow gas at the scrubber overhead

The approach suggested in this section illustrates the dynamic effects that the static PLS or the existing physical model cannot describe. In this study, the PLS model is based on time shifted X block and Y data. The delay range is selected between 0 to 15. In table 3.3 the commutative Y variance (R^2Y) of all models has been summarized.

This analysis shows that the extension of PLS technique to the dynamic model identification yields an improved prediction of product flow gas, when both lagged X and Y blocks are used¹. The prediction result is shown in figure 3.6.

3.4.4 Comparison of PLS and PFPM models in estimating the scrubber top distillation point

The feasibility of identification tool, PLS technique, was examined in the last sections. The case study on the Syncrude Fluid Coker illustrated the construction of a simple but efficient inferential model that can be implemented to perform automatic control of the Coker. Comparing the prediction results from PLS and the pseudo first principles model in estimating the scrubber overhead distillation point, it can be shown that the PFPM provides

¹Due to the reason that the PFPM, presented in the last chapter, is a static based model it is required to apply the calculated time invariant product flow gas instead of dynamic one.

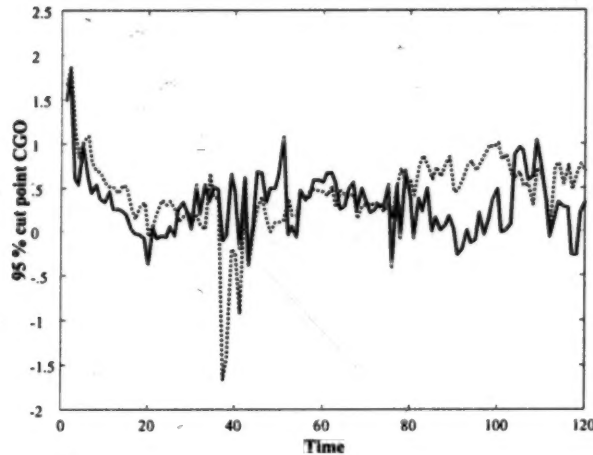


Figure 3.3: Predictions obtained using static PLS based empirical model on the cross validation data set. The solid lines are the output from lab and the dots represent the model predictions.

a slightly better estimation while the PLS algorithm predicts the key quality variable in a simpler model with less variables but the prediction is close enough to PFP based model. This can be confirmed by comparing the mean square error (MSE) from both approaches:

$$MSE_{PLS} = 7.5$$

$$MSE_{PFP} = 5.5$$

As mentioned before, the PLS model could be improved further more if we have more data points and the priori model structure is more accurately known. The prediction results from both techniques are highlighted in figure 3.7

3.5 Introduction to the GMDH technique

It was shown in the previous sections that in many types of mathematical and physical models, to choose the set of polynomial terms, we require the priori knowledge of the system. However, when very little knowledge of the process is available, various assumptions may be made which lead us to a model with less reliability. The concept of group method of data handling or polynomial nets, invented by a Ukrainian cyberneticist, A. G. Ivakhnenko, provides a powerful architecture to form a polynomial function of all or some of the input variables without having specific knowledge of the system or massive amounts of data.

The GMDH combines both statistics and Neural Networks to discover a model structure automatically. The network is implemented with polynomial terms in the links and a genetic component to decide how many layers need to be built. The result of training at the output layer can be represented as a polynomial function of all or some of the inputs. Providing a

Tag- name	Description
T13	Reactor temperature
F8	Bitumen feed to the scrubber
F541	Top feed ring
F542	Middle feed ring
F543	Bottom feed ring
F18	Scrubber pool quench
F10	Fresh Bitumen to the reactor
F5023	VDU bottoms to the reactor
F514	Cocker feed Bitumen plus Residual
QX909	Scrubber overhead heat
T13*F5023 F5023*QX909	NON-LINEAR TERMS

Table 3.2: Process variables for the product flow gas

parsimonious model, the GMDH technique gives an increased insight into the structure of a complex system and the algorithm "optimally" selects (1) what variables appear in which equations, (2) the optimum degree of nonlinearity of the resulting model equations and (3) the structure and degree of interaction among variables. The important feature of the iterative GMDH algorithm is its ability to identify both linear and non-linear polynomial models. The traditional algorithm known as the Ivakhnenko polynomial method is briefly described as follows:

Designating input variables as $x = x_1, x_2, \dots, x_n$, we combine each two variables (x_i, x_j) by the regression equation

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_ix_j$$

In other words, the first layer created is made by computing the quadratic regressions of the input variables and then choosing the best ones. The second layer is created by computing regressions of the values in the first layer along with the input variables. Again, only the best are chosen by the algorithm. These are called survivors. This process continues until the net stops getting better (according to a prespecified selection criterion). This computation procedure has been described in the simple scheme shown in figure 3.8².

The traditional algorithm of the group method of data handling is explained in more detail in the following section.

3.6 GMDH algorithm

A more thorough description of the basic Ivakhnenko polynomial algorithm is provided in this section. This method is based on the sorting-out procedure, testing the models and

²Figure obtained from the book "Self-Organizing Method in Modeling, Farlow"

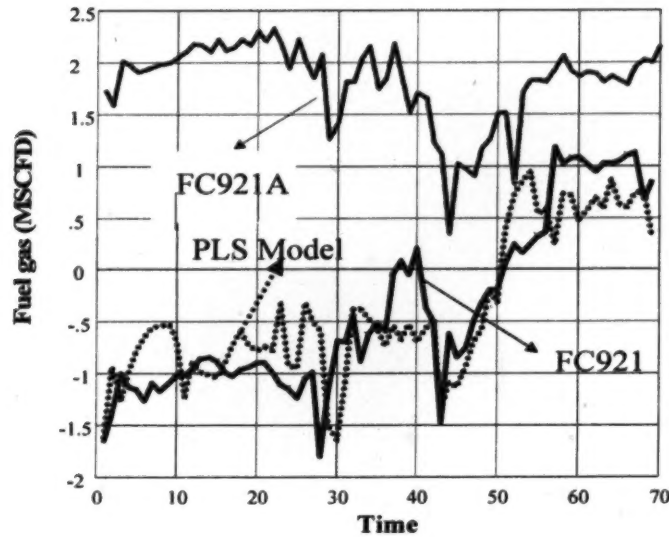


Figure 3.4: Predictions obtained from static PLS based empirical model on the training data set. The solid lines are the actual output, product flow gas (FC921), and the dots represent the model predictions. FC921A represents the existing regression based model.

choosing the best candidates in each generation based on the prespecified criterion. Having n observations, first we subdivide the data points into two sets, training and testing observations (figure 3.10) and follow the steps indicated below.

- Step 1. Among all the input variables, we take two at a time, apply the regression method and construct the quadratic polynomial from the observations y_1, x_{1p}, x_{1q} for different $p, q = 1, \dots, m$ over n data points. These new calculated observations will be stored as new variables in a new array Z . The new calculated variables make a better prediction of y_1 compared to the original generation $x_1, x_2, \dots, x_m$ and will be computed in a similar manner in the next layers. To choose which variables should survive and go to the next generation, we require an objective criterion and this is the stage that the testing data set plays a key roll.
- Step 2. Defining the root mean square called as regularity criterion (r_j), we calculate this value for each new variable Z and order the column of Z according to increasing r_j^3 .

$$r_j^2 = \frac{\sum_{i=nt+1}^n (y_i - z_{ij})^2}{\sum_{i=nt+1}^n y_i^2}$$

³There are different criterion applied in different softwares. We have to emphasize that, here, the original GMDH has been introduced

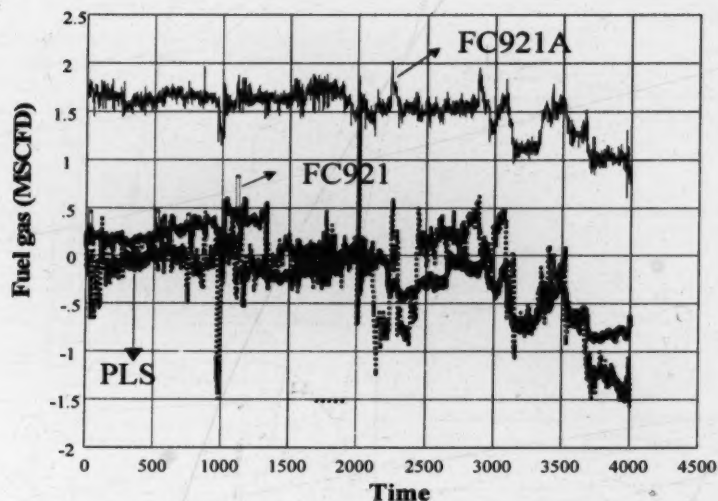


Figure 3.5: Predictions obtained from static PLS based empirical model on the cross validation data set. The solid lines are the actual output, product flow gas, and the dots represent the model predictions. FC921A represents the existing regression based model.

In the Z column, the elements satisfying $r_j < R$ will be selected for the next generation where R is a number, prespecified by the user. The goodness of fit r_j will be summed over the observations in the testing set. Note that the number of new inputs at each layer might be less or greater than the original independent variables.

- Step 3. If the smallest value of the goodness of fit (RMIN) in the current generation is less than RMIN of the previous layer, we repeat step 1 and 2. However, if this value is larger, the algorithm should be stopped. From figure 3.12, it can be seen that after four iterations the process should be stopped and the minimum RMIN in the plot refers to the Ivakhnenko polynomial or final model.

3.7 Application of GMDH in estimating the overhead scrubber distillation point

Using the same input measurements introduced in table 3.1 except the nonlinear terms⁴, we perform the group method of data handling to construct an inferential model for the 95%

⁴Since all these techniques, automatically, provide the nonlinear model, there is no need to add the nonlinear terms used in PLS model. The polynomial complexity can indeed be found by GMDH technique

Model	R^2Y
Static PLS model	0.71
Dynamic PLS model with time shifted X block	0.67
Dynamic PLS model with time shifted X and Y block	0.85

Table 3.3: Commutative Y variance of static and dynamic PLS models.

distillation point CGO. The results and analysis are summarized in the following section.

3.7.1 GMDH analysis

Initializing the output and input variables and selecting advanced GMDH net architecture, we specify a third order nonlinear polynomial as the maximal degree of the model.⁵ The criterion according to which the best models are selected (objective function) is set to the full complexity prediction squared error (FCPSE) which can be defined in the following expression:

$$FCPSE = Norm.MSE + CC * var(a) * C/N \quad (3.23)$$

where, *Norm.MSE* is the average squared error of the model on the training set, *N* is the number of patterns in the pattern file, *var(a)* is the variance of the actual output variable, *CC* is the Criterion Coefficient to change the weight of the over fitting penalty and finally *C* is the overall model complexity coefficient, which takes into account the complexity of each term in the model. Training the 120 observations collected from April to July 1, 2000, we achieve the model in the nonlinear polynomial form. The performance of this approach on the testing data set is shown in figure 3.13.

⁵The algorithm is the same as the original Ivakhnenko method except that the nonlinear regression equation has the form:

$$y = A + Bx_1 + Cx_2 + Dx_1^2 + Ex_2^2 + Fx_1x_2 + Gx_1^3 + Hx_2^3$$

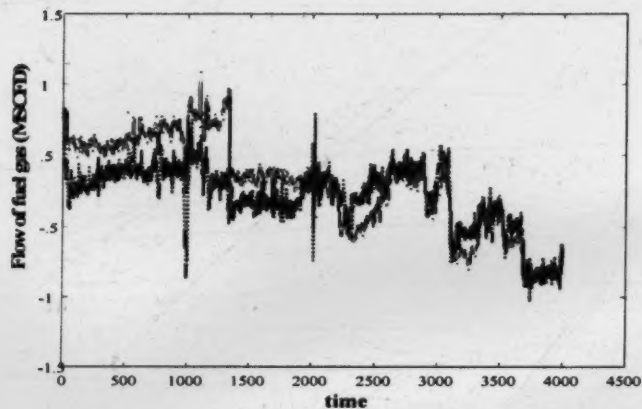


Figure 3.6: Predictions obtained using dynamic PLS based empirical model on cross validation data set of product flow gas. Both X and Y block are time shifted. The solid lines are the actual output and the dots represent the model predictions.

3.7.2 Comparison of GMDH and PLS approach in predicting scrubber overhead distillation point

Comparing the prediction result of the Coker case study using GMDH and PLS techniques (figure 3.14) shows that the GMDH approach provides a better estimation with less mean square prediction error:

$$MSE_{PLS} = 7.4834$$

$$MSE_{GMDH} = 6.3663$$

The GMDH analysis also highlights the advantages of this identification tool in the sense that the model structure is self organized, a better prediction of output variable is achieved and the relative contribution of each input to the output variable is presented.

3.8 Introduction to Neural Networks

Artificial Neural networks (ANN) have been applied in solving a wide variety of problems in science and engineering in the recent years. Among the various applications of ANN such as interpretation, diagnosis, process monitoring, classification and pattern recognition, the concept of nonlinear system identification and prediction are the most common use of neural networks.

The idea of neural networks originated from the human brain, which performs intelligent operations. Like the brain, neural networks are formed from hundreds or thousands of simulated neurons connected together in much the same way as the brain's neurons. Like people, neural networks learn from experience and do not need formulas or rules. They are

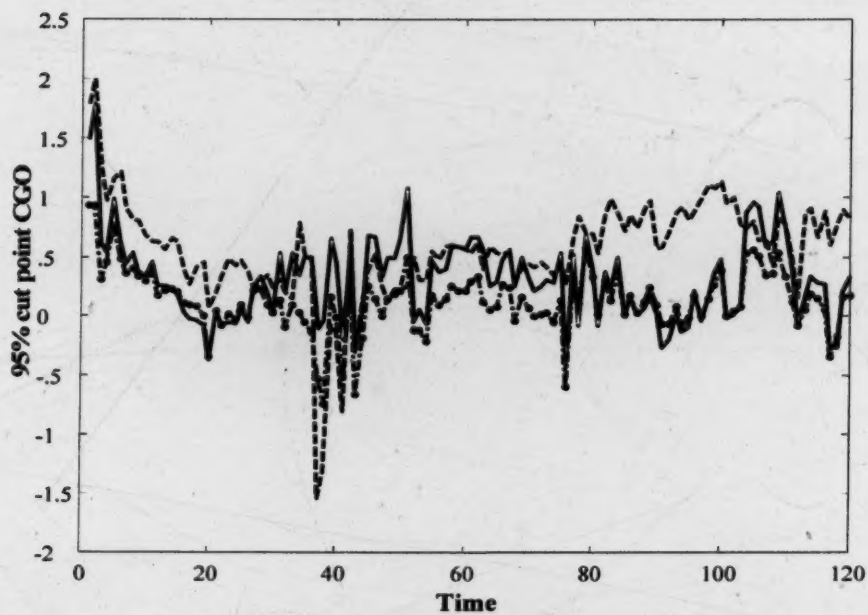


Figure 3.7: Comparison of PLS and PFPM predictions of scrubber overhead distillation point. The solid lines are the actual output. The dots and dashes represent the PFPM and PLS estimations respectively.

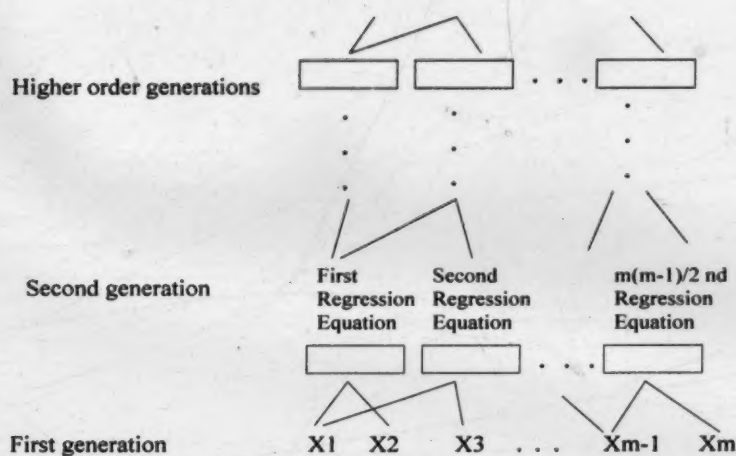


Figure 3.8: Basic scheme of propagations of variables in GMDH approach

trained by repeatedly presenting examples including both inputs and outputs. The network tries to learn each of the examples in turn, calculating its output based on the provided inputs. If the network output doesn't match the target output, it will be corrected by changing its internal connections. This trial-and-error process continues until the network reaches the user specified level of accuracy. Once the network is trained and tested, we can give it new input information, and it will produce a prediction.

We can distinguish neural networks by two different learning algorithms, supervised and unsupervised NN. In supervised learning, as indicated before, the correct results (target values, desired outputs) are known and are given to the NN during training so that the NN can adjust its weights to try match its outputs to the target values. After training, the NN is tested by giving it only input values, not target values, and seeing how close it estimates the outputs to the correct target values. In unsupervised learning, the NN is not provided with the correct results during the training. In other words, unsupervised learning involves no target values. This learning is used in a wide variety of fields under a wide variety of names, the most common of which is "cluster analysis".

Among different types of ANN, two major kinds of network topology are feedforward and feedback. In a feedforward NN, the connections between units do not form cycles. In a feedback or recurrent NN, on the other hand, there are cycles in the connections. Feedback NNs are usually more difficult to train than feedforward NNs.

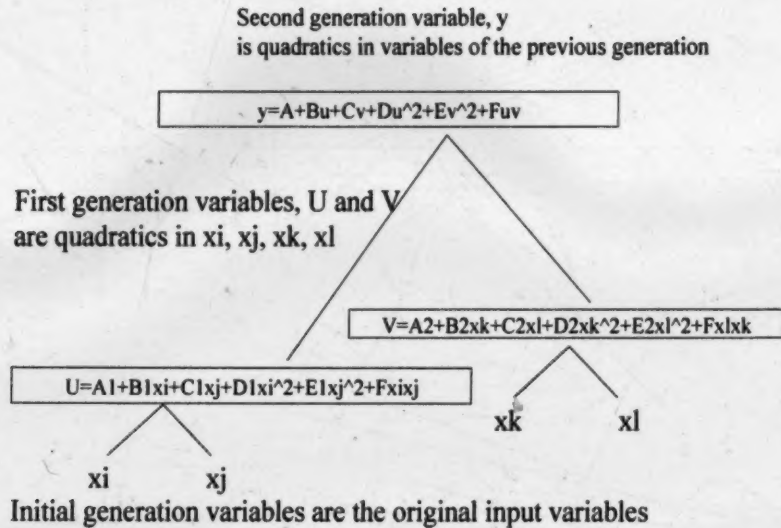


Figure 3.9: Generalization of GMDH algorithm

In the following section, after introducing some fundamentals of NN, we review the basic architecture of neural networks, back propagation algorithm, in more detail.

3.9 Fundamentals of Neural Networks and Back Propagation Algorithm

3.9.1 Simple Neurons and Networks

A very simple neural network can be characterized by its architecture or its pattern of connections between the neurons. A neuron, unit, or node is the basic building block of an artificial neural network, which has some inputs and outputs. Each neuron is connected with other units by directed communication links with an associated weight. Each neuron has also an internal state called its activation function. After each input is weighted by a factor w_i , the activation function f receives the whole sum of weighted inputs. The neuron sends the output from this nonlinear function as a signal to several other neurons. ANN is generally built by putting the neurons in layers. Based on the method of determining the weights on the connections - training or learning algorithm of the network - there are different architectures of NN. Figure 3.15 shows the simple three layers feedforward structure. The first layer distributes the inputs to the second layer called the hidden layer. The last layer is the output layer. Each output unit computes its activation to form the

Y (output variable)		X (Input variables)			
Y1		x1	x2		xm
Training observations	y1	x11	x12	.	x1m
	y2	x21	x22	.	x2m
	.	.	.	.	.
	.	.	.	.	.
	ynt	xnt,1	xnt,2	.	xnt,m
Testing observations	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	yn	xn1	xn2	.	xnm

Figure 3.10: Input to the GMDH algorithm (one output Y1 and m independent variables).

response of the net for the given input pattern.

The general algorithm of feed-forward neural networks trained by back propagation is discussed in the following section.

3.9.2 Back Propagation Algorithm

Referring to figure 3.16, the back propagation algorithm on a three layer feedforward network functions as follows:

Setting small random numbers, we initialize the weight values. Each neuron in the hidden layer receives a signal from the neurons in the previous layer (input layer), and each of those signals is multiplied by its initialized weight value. The weighted inputs are summed and passed through a limiting function which scales the output to a fixed range of values.

Two common activation functions, typically used in ANN are:

- sigmoid function, $f_1(x) = \frac{1}{1+\exp(-x)}$
- bipolar sigmoid function, $f_2(x) = \frac{2}{1+\exp(-x)} - 1$

The output of the limiter is then sent to all of the neurons in the next layer (output layer). After propagating the signals through the network and reading the output or target value, we need a method to adjust the weight values. An applied, common learning algorithm used in solving most problems is back propagation (BP). Providing a learning set of input and known-correct output, a BP learns what type of behavior is expected and it tries to adapt the network.

The BP learning process works in small iterative steps: After producing the output based

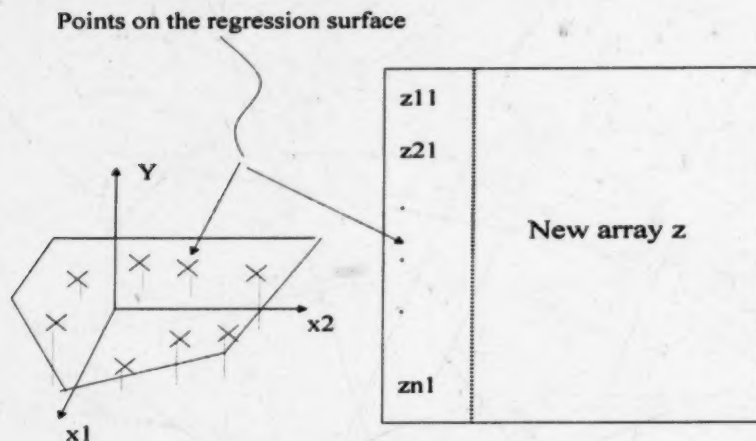


Figure 3.11: Construction of the new array Z.

on the initialized weighting values indicated above, this output is compared to the known-correct output and a mean-squared error signal is calculated. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal.

The whole process is repeated from the first step. The cycle is repeated until the overall error value drops below some pre-determined threshold. At this point we say that the network has learned the problem "well enough". Note that the network will never exactly learn the ideal function, but rather it will asymptotically approach the ideal function.

3.9.3 Neural Network analysis

We apply the standard type of feedforward neural network using the back-propagation learning algorithm to the nonlinear system Fluid Coker. The training and validation data set are the same as used in GMDH and PLS techniques. The prediction performance of neural networks on the training and testing data sets are shown in figures 3.17 and 3.18.

3.9.4 Comparison of prediction result from GMDH and Neural networks algorithms

In the last two sections we inferred the variable 95% cut point CGO employing GMDH and NNs. The prediction result obtained from both algorithms is shown in figure 3.19. Comparison the mean square error shows that NNs provide a close estimation to GMDH.

$$MSE_{NNs} = 6.8456$$

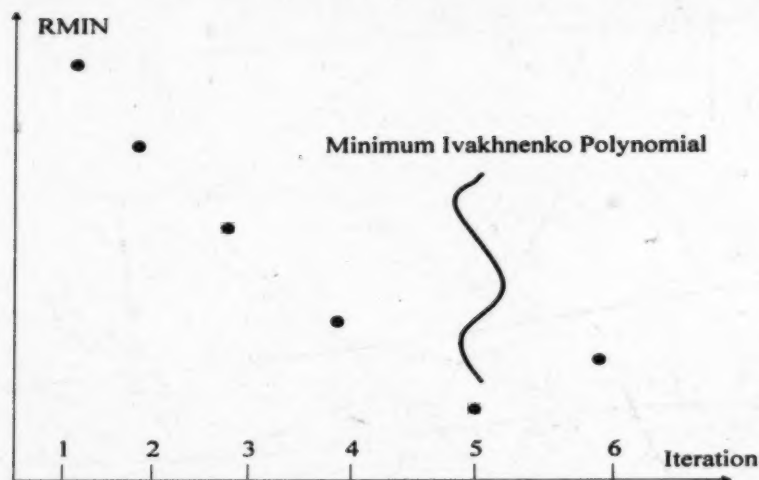


Figure 3.12: Stopping criterion.

3.10 Conclusion

In this work, different identification algorithms for nonlinear systems have been presented. These tools have been applied on the industrial Fluid Coker to infer the variable 95% cut point CGO. It was shown that among all black box techniques, Group Method of Data Handling provides a better estimation while the model structure is self-organized. The prediction result from PLS technique, on the other hand, revealed that the feasibility of this approach depends on the availability of the model structure under consideration. If such kind of prior information is easily at hand, PLS gives a very close estimation to GMDH. Neural Networks provided reasonably successful prediction, which was very close to GMDH estimation. However, the main gap between this mathematical tool and the practical application is that there is no explicit model available and this tool usually requires more data points.

To evaluate the feasibility of all techniques compared to the Psuedo First Principles modeling, we compared the prediction error of all these approaches. The presented models could be ranked based on the prediction result on *this case study* in an descending order as follows:

- 1. PFPM
- 2. GMDH
- 3. NNs
- 4. PLS

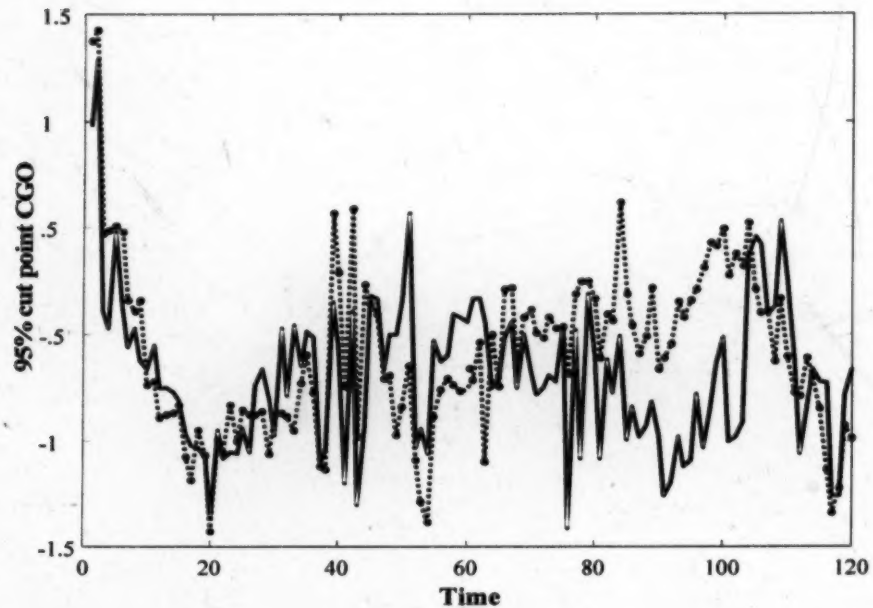


Figure 3.13: Prediction of 95% cut point CGO using group method of data handling empirical model on the testing data set. The solid lines are the output from lab and the dots represent the model predictions.

MODEL	FPM	PLS	GMDH	NNs
Structure-selection	Model structure is found upon the physical law	Unable to search model structure	Model structure is Searched automatically	No explicit model structure is available
Application	Applicable to the system with exact knowledge of the process (linear/nonlinear) Requires more cost and time	Mostly for linear model but can also be used for nonlinear model	For both linear and nonlinear systems	Prediction for both linear and nonlinear processes
Prediction performance	Exact prediction is achieved if the priori knowledge of the process is available	Depends on the availability of data points and model structure	Provides a good prediction if the model structure is restricted to a polynomial form	Large sampling size is a definite requirement

Table 3.4: Comparison of the black box techniques with first principles model

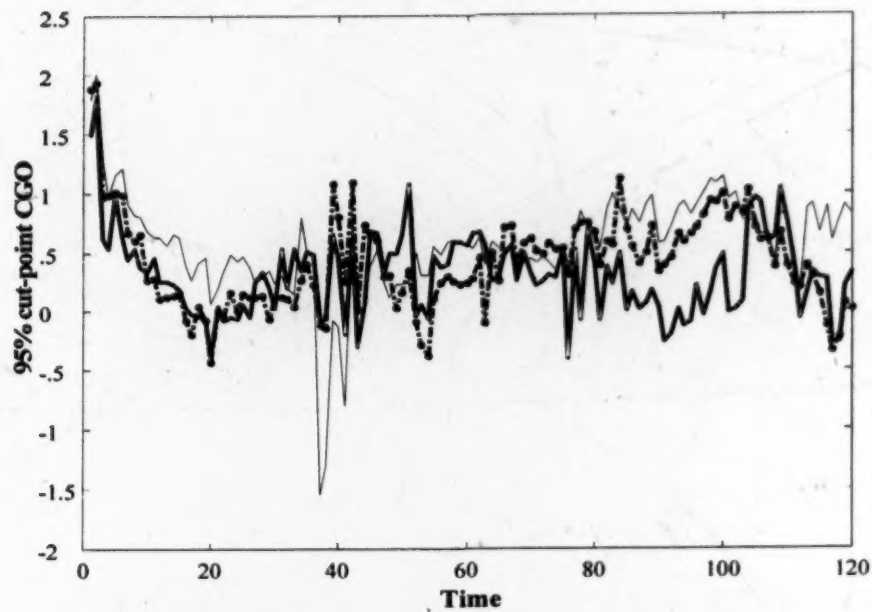


Figure 3.14: Comparison of GMDH and PLS prediction result in estimating the 95% cut point CGO. The solid lines and dots are the actual output from lab and the model predictions from GMDH respectively. The dashes represent PLS estimation.

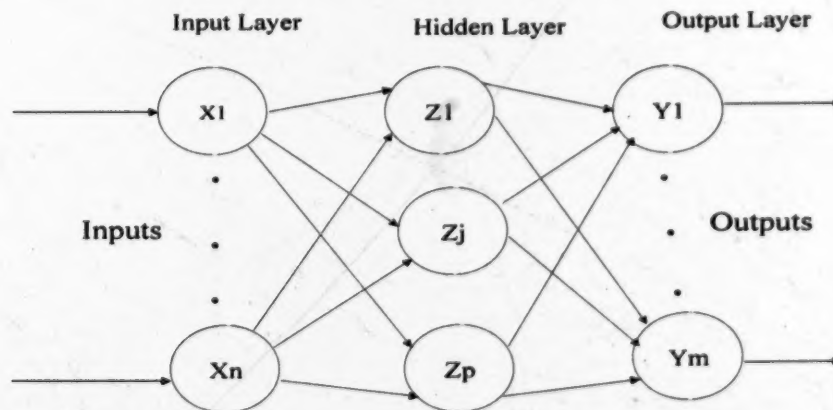


Figure 3.15: Feedforward neural network with one hidden layer.

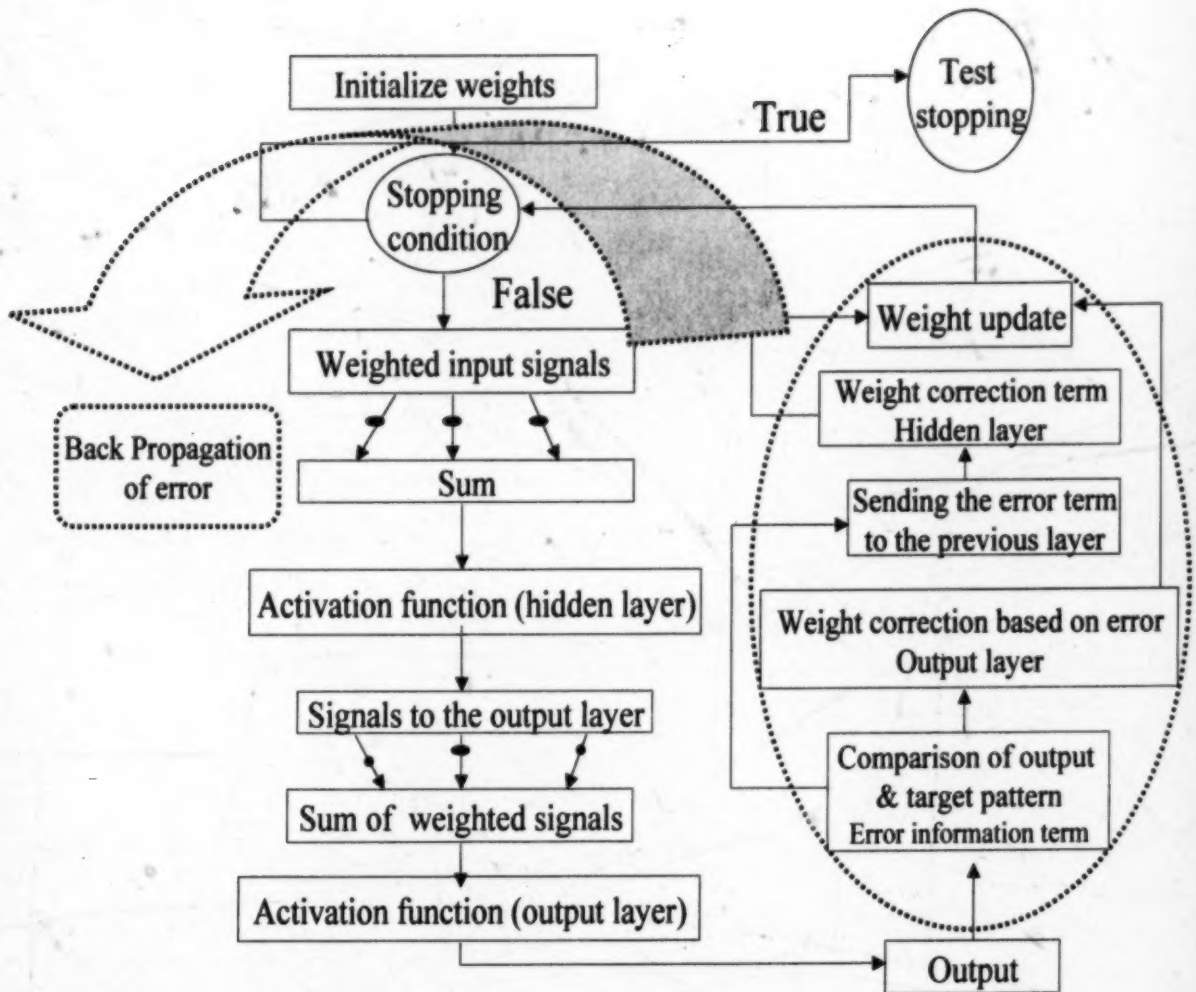


Figure 3.16: Flow chart of back propagation algorithm in three layer feedforward artificial neural networks.

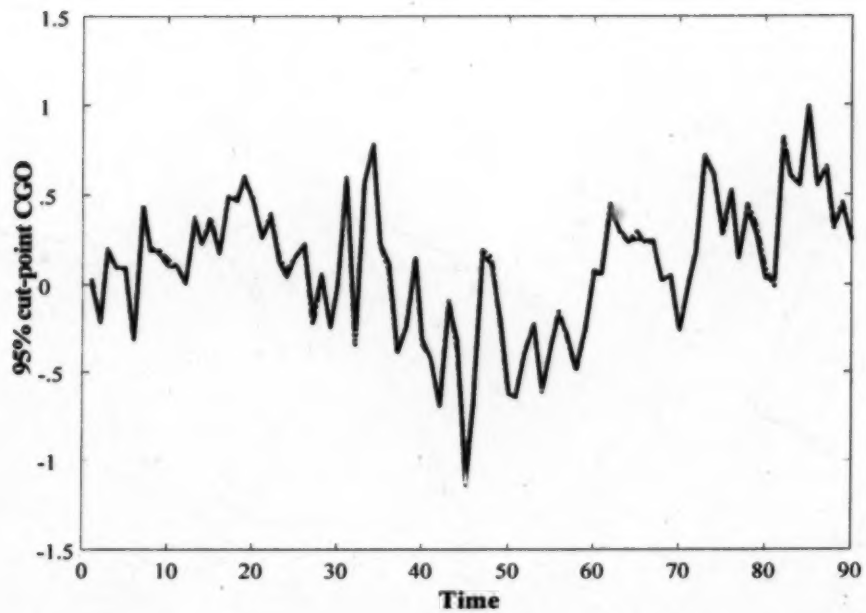


Figure 3.17: Prediction of 95% cut point CGO using feedforward neural networks with back propagation learning algorithm on the training set of data. The solid lines are the actual output and the dots represent the model predictions.

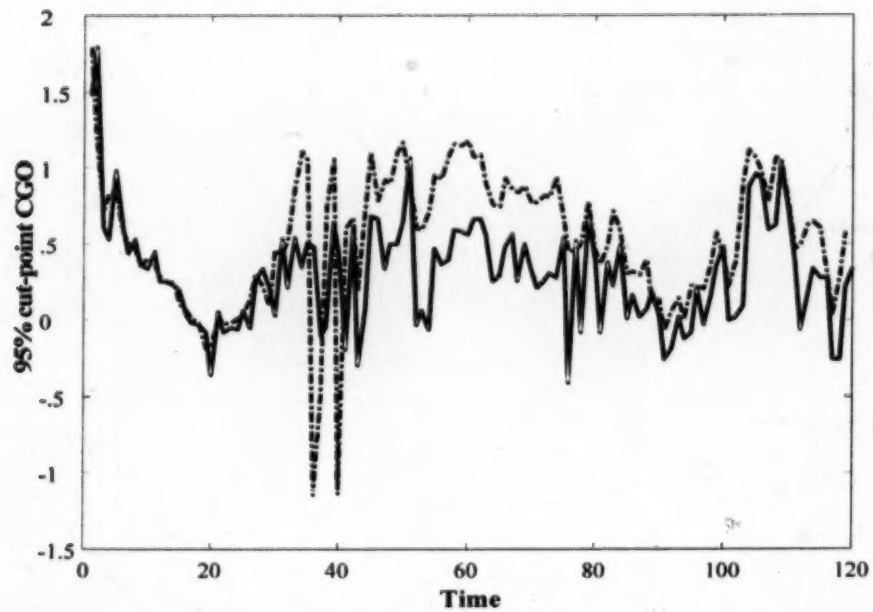


Figure 3.18: Prediction of 95% cut point CGO using feedforward neural networks with back propagation learning algorithm on the testing set of data. The solid lines are the actual output and the dots represent the model predictions.

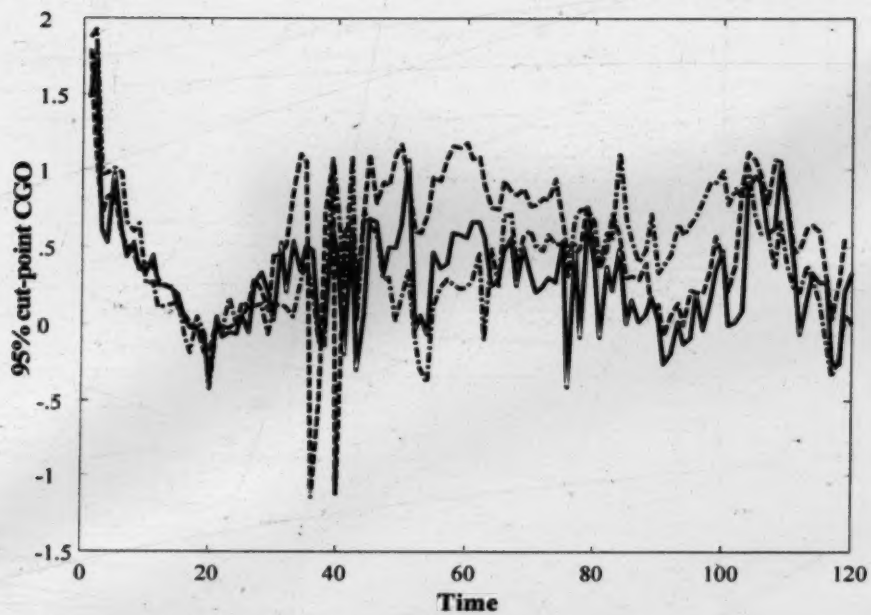


Figure 3.19: Prediction of 95% cut point CGO using NNs and GMDH. The solid lines are the actual output, the dots represent the model predictions from GMDH and the dashes are the estimation from NNs.

Chapter 4

Issues in Experimental Design for Nonlinear System Identification

4.1 Abstract

The objective of this chapter is to study the practical aspects of nonlinear system identification including the design of input signals capturing the underlying dynamical behavior of the NARMAX model. We demonstrate the feasibility of developing a nonlinear model using GMDH, nonlinear regression and Genetic Algorithms through experiments on a CSTR pilot scale reactor. The practical goal is to provide a nonlinear inferential model to analyze the concentration of the key component in a second order chemical reaction.

4.2 Introduction

Analysis of the input/output behavior of a physical system and constructing a model, which captures the key underlying dynamics of the process, is the first step in controlling a system.

Since most industrial plants are complex and exhibit a certain degree of nonlinearity, usually, the global behavior of the system over the whole operating range can be described by nonlinear models rather than by linear ones that are only able to approximate the system around a given operating point. A popular and general class of nonlinear models, discussed in this chapter, is the one corresponding to the so-called NARMAX (Nonlinear Autoregressive Moving Average models with eXogenous inputs) which consists of nonlinearity in both input and output. Various nonlinear models have been discussed in detail by Haber and Unbehauen (1990).

When little process information is available and the structure of the model is not given *a priori*, developing a "robust" nonlinear model usually requires two phases:

- Structural identification in which the general form of the equations that govern the unknown dynamics are determined. The structure identification is usually organized according to the classes of nonlinear dynamical models and to the kind of experiment performed on the unknown process[23].

- Parameter estimation in which coefficient values that match the model are found.

Designing an input sequence has particular significance in the development of a nonlinear model and identification of both parameters and structure of the system. Only an appropriate frequency rich input excitation can expose the underlying dynamics of the system. Among the different types of the excitation signals, one of the most widely used is the pseudo-random binary sequence (PRBS). However, due to the bi-level of the 'RBS' signal in general, they may not be desirable for nonlinear systems. Characterizing the 'RBS' input signal as $u(k) = \pm a$ where a is the amplitude of the binary signal shows that the negative terms may be eliminated from the series if any term of the model has an even power. A detailed review of the theory and application of RBS signal is provided in Godfrey (1996). Three and five level sequences are other types of input signals used in the identification of nonlinear systems. In (Barker et al., 1972) these signals are considered for identification of the Volterra kernels. A review of the three level sequences is presented in (Barker and Davy, 1978).

In this chapter, after providing a short description of a pilot scale CSTR reactor, the practical issues of the identification of a nonlinear system are investigated. We compare different classes of input signals including sum of sine waves and uniformly distributed multi-level sequences through both simulations and experiments and propose a modified multilevel signal. Other important issues in designing the input signals, namely, frequency and amplitude of the excitation signals are also addressed in this work.

For the continuous stirred tank reactor, the physical model has been developed in the continuous time domain. The general nonlinear discrete time model shows that a NARMAX model describes the dynamical behavior of the physical system.

A very powerful identification tool, Group Method of Data Handling, has been discussed in the previous chapter. This technique combines the two phases of system identification and provides a useful algorithm for modeling complex nonlinear systems that have a large number of variables and parameters but a relatively small amount of collected data[25]. We apply this approach to identify the parameters as well as model structure of the CSTR reactor.

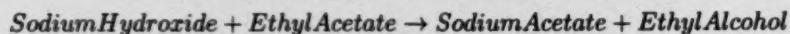
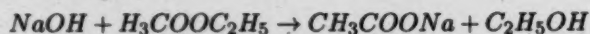
Regression analysis is applied as another tool to identify the model parameters, assuming the structure of the model follows the mechanistic equation. The performance of the identified models is discussed in detail in each section.

4.3 Pilot Scale Continuous Stirred Tank Reactor

The process studied in this chapter is a CSTR reactor designed to demonstrate the mechanism of a chemical reaction. Reaction is monitored by conductivity probe as the conductivity of the reacting solution changes with conversion of the reactants. This provides a convenient method for monitoring the progress of the reaction instead of using the inaccurate process

of titration.

The reaction chosen is an equi-molar, overall second order chemical reaction between ethyl acetate and sodium hydroxide:



By an overflow tube the volume of the reactor is fixed at 1.75 L. To maintain the reaction at a precise temperature, an on/off local controller is used. Heating the water passing through the coils wind, the temperature is held steady at the desired set point. The concentrations of the reactants are both held at 0.1mol/L . The reactants are attached to the separate pulsating pumps feed into the reactor. The schematic diagram of the CSTR reactor is shown in figure 4.1.

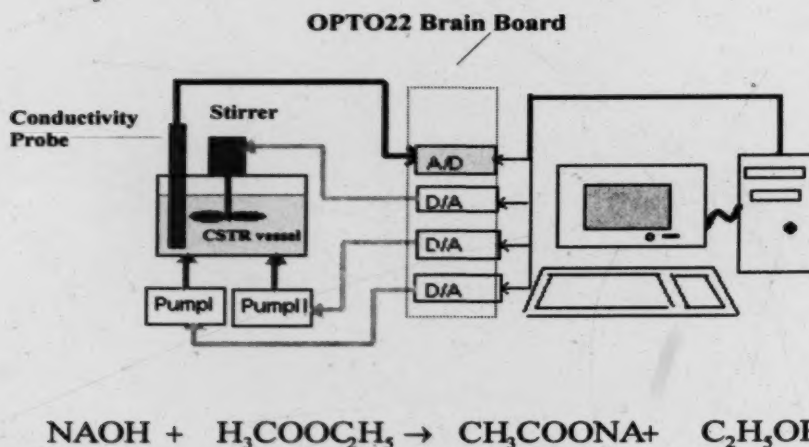


Figure 4.1: Schematic diagram of CSTR reactor

For identification purpose, the conductivity of the product in the reactor at time t , λ , is considered as the output variable of interest. The flow rates of the reagents, Sodium Hydroxide (F_a) and Ethyl Acetate (F_b) are the manipulated and disturbance variables limited by the maximum and minimum speeds of the pumps for a range of 0-80 mL/min. Assuming the solution in the reactor is well mixed, the speed of the agitator is set at its maximum value, 175 rpm. The CSTR apparatus is controlled remotely using an OPTO22 brain board for D/A and A/D outputs and inputs in LABVIEW. The configuration of the reactor and OPTO22 is set up with 3 output channels, pump speeds on the two feed pumps and agitator speed, and one input channel, conductivity of the solution.

To achieve an initial understanding of the dynamics of the system, we develop the mechanistic equation based on the physical laws relating to the system. The overall mass balance can be stated as follows:

$$\text{Accumulation} = \text{Input} - \text{Output} + \text{Generation}$$

Using the balance equation for the reactant a in the reactor with volume V^1 , we have:

$$\frac{d(Va_1)}{dt} = F.a_0 - F.a_1 - V.k.a_1^2 \quad (4.1)$$

Conversion of the continuous-time differential equation to the discrete time domain with the sampling rate T_s shows that the system has nonlinearity in a cross term of input-output and a square term in the output variable a_1 .

$$a_1(t) = \beta_1 F_a(t-1) + \beta_2 a_1(t-1) + \beta_3 a_1(t-1)(F_a(t-1) + F_b(t-1)) + \beta_4 a_1^2(t-1) \quad (4.2)$$

It should be mentioned that the degree of conversion of the reagents affects the conductivity of the reactor contents. Thus, the conductivity recorded by the data logger is proportional to the concentration and can be used to infer the concentration of Sodium Hydroxide/Sodium acetate and therefore the amount of conversion. As such, the physical model for conductivity follows the equation 4.2 and has the same structure as the balance equation.

$$\Lambda(t) = z_1 F_a(t-1) + z_2 \Lambda(t-1) + z_3 \Lambda(t-1)(F_a(t-1) + F_b(t-1)) + z_4 \Lambda^2(t-1) \quad (4.3)$$

To determine the model for the output variable λ in the presence of various types of disturbances and to review the effect of different input sequences on the identification of the system, we require a number of experiments under consistent conditions which cannot be easily achieved by experiments on a real system. Due to this reason, the general understanding of the design of the input signals is first verified using the data collected from the simulated process based on the physical model 4.1. Then, we identify the model using GMDH, Linear Regression and GAs tools. Once the simulation yields satisfactory results, the designed input sequence will be applied to the pilot-scale reactor.

4.4 Designing the Input Signal

The first important issue in process identification is to choose an appropriate input excitation signal, which is dependent on the different classes of the models. It is known that a linear time-invariant system with n unknown parameters can be identified by using an input which is a linear combination of sinusoids with $\frac{n}{2}$ distinct frequency components (e.g. Dasgupta et al. 1991), [44]. Another standard signal noted by Leontaritis and Billings (1987) is binary sequences, which are typically useful for identification of linear systems. However, there is little formal theory for input sequence design for nonlinear systems especially for the general class, NARMAX models, described by billings and Voon as follows [2]:

$$y(k) = F(y(k-1), y(k-2), \dots, y(k-p), \\ u(k), u(k-1), \dots, u(k-q),$$

¹See the Nomenclature list in appendix A

$$e(k-1), e(k-2), \dots, e(k-r)) + e(k)$$

where, $F(\cdot)$ is a nonlinear function indicating that both input $u(k)$ and output sequence $y(k)$ are nonlinear and $e(k)$ is a "Gaussian white noise" sequence.

It is clear that applying binary signals may cause loss of information and identifiability for nonlinear systems. The inverse repeat of these signals might eliminate the negative terms from the series if the system contains an even order nonlinearity. Leontaritis and Billings (1987) suggest the use of an input sequence which is uniformly distributed and has an amplitude constraint might be an appropriate signal for the identification of a nonlinear system. Pottmann, *et.al.* (1993) use the equivalent signal on a PH neutralization process with a geometric reason as follows:

They interpret the identification of nonlinear systems with $y(k-1), \dots, y(k-m), u(k-1), \dots, u(k-n)$ independent variables as an approximation problem in R^{n+m} . Only the surface which is well fitted to the available data in R^{n+m} will describe the dynamical behavior of the system. To achieve this goal, the data points should be well distributed in the domain of R^{n+m} corresponding to the normal operating conditions of the process [44]. A surface which is based on the uniform distribution of the data points switching between the maximum and minimum values yields an accurate description of the process [5]. Such a signal follows the recommendation of Leontaritis and Billings, for the case of amplitude constraint on the input signal. However, the considerable valve movement during experimentation is one drawback of using pure random signals. Due to this reason, Pottmann, Unbehauen and Seborg (1993) suggest applying only small input variations over different operating points of the process.

Through extensive simulation studies it is shown that the above strategy in designing the input signal gives promising results for identification of a NARMAX model obtained from a CSTR pilot scale reactor. The experimental and simulation results of using different excitation signals are discussed in more detail in the following section.

4.5 Illustrative Example In Issues of Excitation of the Input Sequence

The frequency for which the signal should switch values is one important parameter in designing the input sequences, delivering the most useful information in the dynamical behavior of the system. To achieve an initial understanding of the frequency of the designed input for the CSTR reactor, the response of the step input has been studied. It shows that the system needs about 1000s to reach the 63% of the final steady state value. Therefore, the behavior of the dynamical system is relatively slow and the designed input should change with a compatible rate.

The calculation of the frequency of designed signal is better illustrated using a Bode diagram shown in figure 4.2. As it is shown, 70.7% of the maximum amplitude of the excitation signal gives an approximate bandwidth of the process. To ensure that the entire

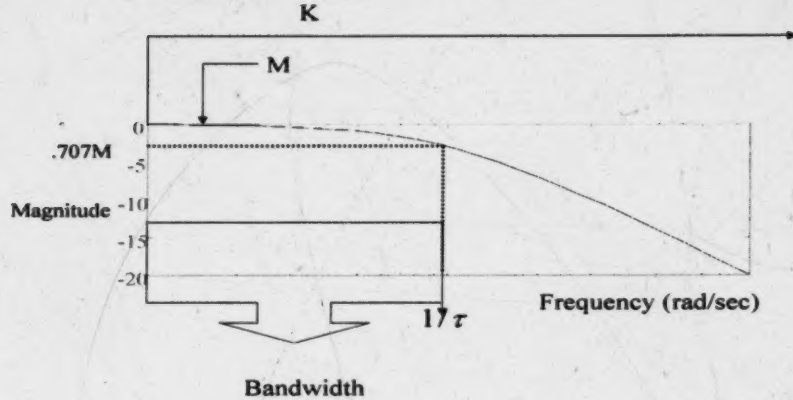


Figure 4.2: Illustration of frequency calculation using a Bode diagram

band spectrum of interest is included in the designed input sequence, a constant $K > 1$ is used to scale the calculated bandwidth. Using the notation of Nyquist frequency of MATLAB, the desired frequency of the excitation signal can be calculated as follows:

$$f = \frac{KT_s}{\tau\pi} \quad (4.4)$$

where T_s is the sampling time, τ is the time constant (found from the step test indicated before) and K is a constant term with a typical range of 2 – 3. This information is entered into the *idinput* command in MATLAB to design the input signal.

Using this information as an approximate guideline for the nonlinear CSTR process, we perform following experiments and simulation around different operating points of the process.

4.5.1 Experiment I: Applying Sum of Sine Waves as an Input

A data set containing 2200 input-output samples with the sampling period 10s were generated from the pilot-scale reactor. The system was considered as the SISO system with the flow rate of Sodium Hydroxide as the input (F_a) and the conductivity as the output variable (λ). The lower and upper limits of the excitation signal were set between 10 and 40mL/min. Applying GMDH tool, a linear polynomial model is found. The results of this identification are summarized in table 4.1 :

Identified model using GMDH tool:

$$\lambda(t) = -.1327 + \lambda(t-1) + 0.009 * F_a(t-1) \quad (4.5)$$

Input Signal	Sum of sine wave
Amplitude of excitation signal	10-40 Ml/min
Sampling time	10
Best formula (GMDH)	$Y=3.3E-003+X2+9.E-003*X1$
Legends	$X1=2.*(Fa(t-1)-13.75)/31.6-1.$ $X2=2.*(\lambda(t-1) -11.06)/3.17-1.$ $Y=2.*(\lambda(t) -11.06)/3.17-1.$
Linear Regression	$\lambda(t) = 0.9806 * \lambda(t-1) + 0.0107 * F_a(t-1) * \lambda(t-1)$

Table 4.1: Summary of identification results for the CSTR reactor

Identified model using linear regression:

$$\Lambda(t) = 0.9806 * \Lambda(t-1) + 0.0107 * F_a(t-1) * \Lambda(t-1) \quad (4.6)$$

Equations 4.5 and 4.6 show the inadequacy of the identified models to characterize the real nonlinear behavior of the system. Possible reasons to the deviation of the identified models from mechanistic equation can be summarized as follows:

- The input sequence is not frequency rich enough to provide an adequate model of the process. Therefore, special consideration needs to be given for designing the input signal for nonlinear CSTR system.
- The upper and lower input variations may not be large enough to approximate the global nonlinear behavior of the system.
- Due to the fast sampling rate, the disturbance effect is significant.
- The approximate time constant and slow response of the system show that the data should switch the upper and lower bound with a slow frequency. An input signal with a relatively fast frequency makes the signal change its value before getting the appropriate response in the previous instant.

The above discussed issues are examined in the following simulation and experiments.

4.5.2 Simulation I: Applying Multi-Level Input Signal

Following the fact that only an appropriate excitation signal can expose the real dynamical behavior of the system, we apply the strategy presented in [44] and divide the whole

operating range of NaOH flow rate into small intervals and create two sets of four-level uniformly distributed signals known as low-level sequence changing the value at 60, 63, 66 and 69 mL/min and high-level signal switching at 70, 73, 76 and 80 mL/min . According to the frequency information generated from the step test, we produce an RBS sequence between -1 to 1. Then the value -1 is uniformly randomly subdivided into the four low-level values, and the value 1 is uniformly randomly subdivided into the four high-level values. An eight level signal is produced. This signal is uniformly distributed and change its magnitude between 60 to 80 mL/min .

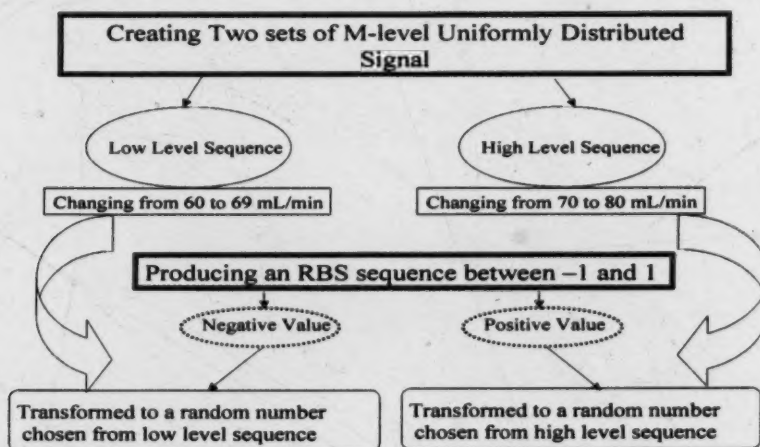


Figure 4.3: Modified Uniformly Distributed M-level Sequence

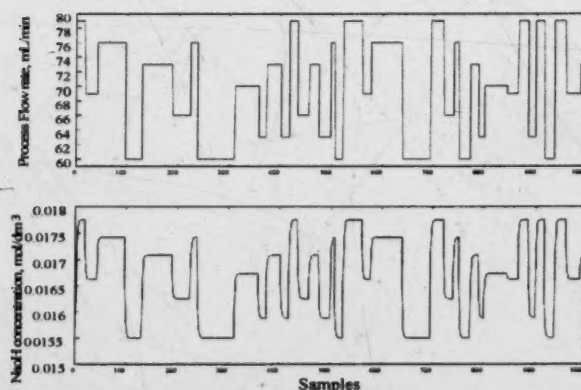


Figure 4.4: Process data for the SISO identification. The uniformly distributed multi-level input signal changes its value between maximum and minimum rate, 60-80 mL/min

The training data sets obtained from simulation with the sampling interval 100s are

Input Signal	8 level uniformly distributed signal
Amplitude of excitation signal	60-80 Ml/min
Sampling time	100
Best formula (GMDH)	$a_1(t) = 0.4125 * F_a(t-1) + 0.565 * a(t-1) + 0.00652$
Linear Regression (LR)	$a_1(t) = 0.5406 * F_a(t-1) + 0.9484 * a(t-1) - 0.1892 * a(t-1) * F_a(t-1) - 0.2494 * a^2(t-1)$
Standard deviation of the identified coeff. by LR	$F_a(t-1): 8.4464e-004$ $a(t-1): 0.0012$ $F_a(t-1) * a(t-1): 0.0012$ $a^2(t-1): 0.0013$

Table 4.2: Summary of simulation results for the CSTR reactor, using an eight level uniformly distributed signal

shown in figure 4.4. Table 4.2 summarizes the identification result. The comparison of the actual and predicted output is presented in figure 4.5.

Identified model using GMDH tool

$$a_1(t) = 0.4125 * F_a(t-1) + 0.565 * a_1(t-1) + 0.00652 \quad (4.7)$$

Identified model using Linear Regression

$$a_1(t) = 0.54 * F_a(t-1) + 0.95 * a_1(t-1) - 0.19 * F_a(t-1) * a_1(t-1) - 0.25 * a_1(t-1)^2 \quad (4.8)$$

The simulation result shows that applying a larger magnitude of the excitation signal and operating the experiment in a higher flow rate of NaoH increases the nonlinearity of the system. Furthermore, employing an information rich data increases the accuracy of the identified model and leads us to a polynomial close to the mechanistic model.

Combination of balance equation and Linear Regression provides an adequate model, which gives considerably good estimate for the concentration of the sodium Hydroxide. Comparison of physical and GMDH identified models shows that the lagged linear input and output variables are selected as prime candidates to predict the concentration of NaoH. Due to this reason, this technique is not capable of identifying the nonlinear terms when their contributions are not considerably significant.

One fact that comes to light from the last simulation is the selection of the sampling rate. As it will be shown in the following section, designing the input signal based on 50s sampling rate shows that the identified model is highly correlated to the lagged output

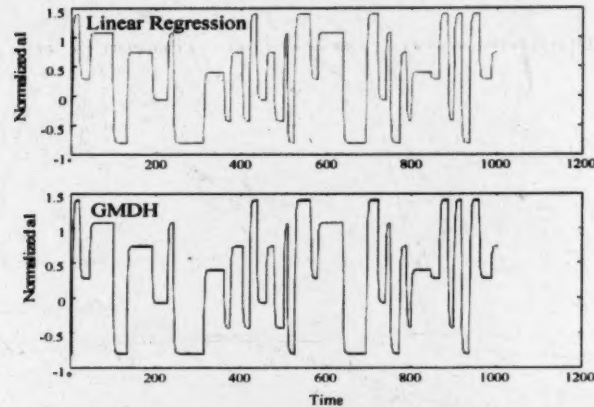


Figure 4.5: Prediction results obtained from Linear regression and GMDH tool using an 8-level uniformly distributed input signal. The solid lines are the simulated output from the CSTR and the dots represent the model prediction.

variable and since this variable plays a relatively important role in predicting the output variable, both LR and GMDH algorithms identify the model in which the corresponding coefficient to the lagged output variable is a dominant term and the rest coefficients are relatively small.

As it is shown in figure 4.6, increasing the sampling time to 200s on the other hand, causes loss of information and provides a poor estimation of output variable which deviates from the actual data. The results are shown in table 4.3 and 4.4.

Normalized Data	Sampling time	Linear regression	GMDH
Lagged-feed	50	0.3088	0.243
Lagged-feed error	//	1.7868e-004	
Lagged-output	//	0.9988	1.13
Lagged-output error	//	2.5672e-004	
Cross-term	//	-0.0980	NA
Cross-term error	//	2.6350e-004	
Square-term	//	-0.1685	NA
Square-term error	//	2.9971e-004	
Constant term	//	NA	-0.00294

Table 4.3: Summary of simulation results for the CSTR reactor, changing the sampling time to 50 s.

Normalized Data	Sampling time	Linear regression	GMDH
Lagged-feed	200	0.6128	0.596
Lagged-feed error	//	0.2385	
Lagged-output	//	1.3133	0.287
Lagged-output error	//	0.2730	
Cross-term	//	-0.0207	NA
Cross term error	//	0.3449	
Square-term	//	-0.7856	NA
Square term error	//	0.3034	
Constant-term	//	NA	0.025

Table 4.4: Summary of simulation results for the CSTR reactor, changing the sampling time to 100 s.

4.5.3 Experiment II: Applying Multi-level Input Sequence to the Real Process

Following the simulation results indicated above, we choose the recommended input signal with the sampling interval approximately equal to $\tau/10$ or 100s and perform the actual identification experiment. Figure 4.7 shows the input-output data after reaching the steady state. The comparison of the data obtained from simulation and real experiment is shown in figure 4.8.

Assuming no fundamental process knowledge is presented, empirical modeling technique, GMDH, is employed to identify the parameters and model structure. As it is shown in table 4.5, this technique is able to identify all linear and non-linear terms presented in the balance equation, i.e. all nonlinear terms are statistically significant. However, the dominant term in the constructed model is the output lagged variable. This result is consistent with the previous simulation results.

Combining the process knowledge and black box technique "Linear Regression", a Semi-Empirical model is provided and final equation describing the dynamical behavior of the CSTR reactor is obtained ².

Identified "LR" Semi-Empirical Model Based on the Normalized Data

$$\Lambda(t) = 0.89 * \Lambda(t-1) + 0.092 * F_a(t-1) - 0.12 * F_a(t-1) * \Lambda(t-1) + 0.16 * \Lambda(t-1)^2 \quad (4.9)$$

Figure 4.7 shows the estimated values of conductivity versus actual data using the model 4.9, presented above.

²The identified model shown in table 4.6 is based on the normalized data.

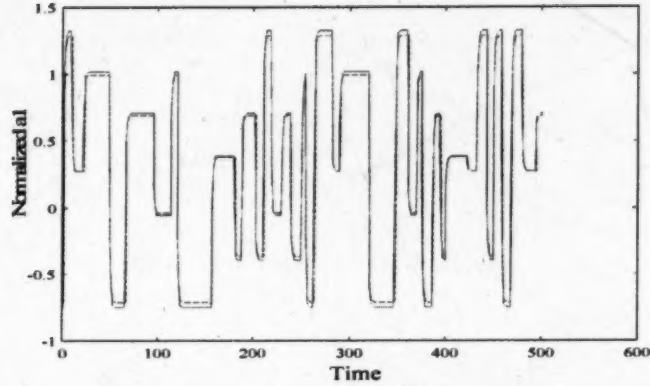


Figure 4.6: Prediction results obtained from Linear regression using the sampling interval equal to 200 s. The solid lines are the output from the simulation and the dash line represents the model prediction.

The cross validation run shown in figure 4.8 indicates that the proposed model captures the dynamics of the system reasonably well. It should be mentioned that the cross validation data is obtained from a different identification experiment. The operating points at this run, are set at $F_{as}, F_{bs} = 10$ mL/min and the temperature of the reactor is kept constant at 30° .

The final dynamical model for the pilot scale CSTR reactor can be presented as follows:

Non-linear Polynomial Model for CSTR System Based on Normalized Data

$$\Lambda(t) = 0.89 * \Lambda(t-1) + 0.092 * F_a(t-1) - 0.12 * F_a(t-1) * \Lambda(t-1) + 0.16 * \Lambda(t-1)^2 \quad (4.10)$$

Non-linear Polynomial Model for CSTR System Based on the Actual Data

$$\begin{aligned} \Lambda(t) = & 6.6 + 0.8962 * (\Lambda(t-1) - 6.59) + 0.0074 * (F_a(t-1) - 69.3) \\ & - 0.0168 * (F_a(t-1) - 69.3) * (\Lambda(t-1) - 6.59) + 0.2716 * (\Lambda(t-1) - 6.59)^2 \end{aligned} \quad (4.11)$$

4.6 Estimation of Conductivity in the Continuous Stirred Tank Reactor, Performing Genetic Algorithm

As mentioned earlier, the objective of this case study is to provide the dynamical model to estimate the amount of conductivity in the CSTR reactor using measurements of inlet flow. We employ Genetic Algorithms (GAs) as another identification tool to estimate the

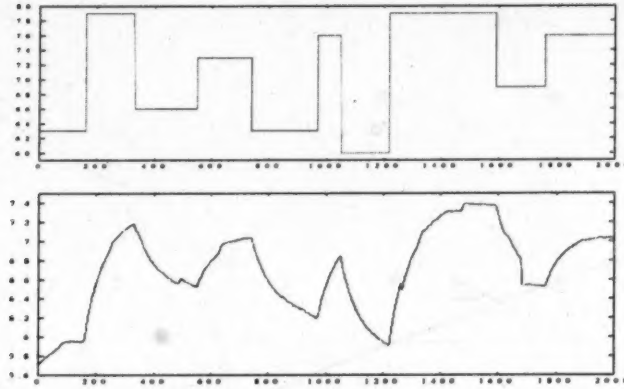


Figure 4.7: Plant Data for identification of CSTR process.

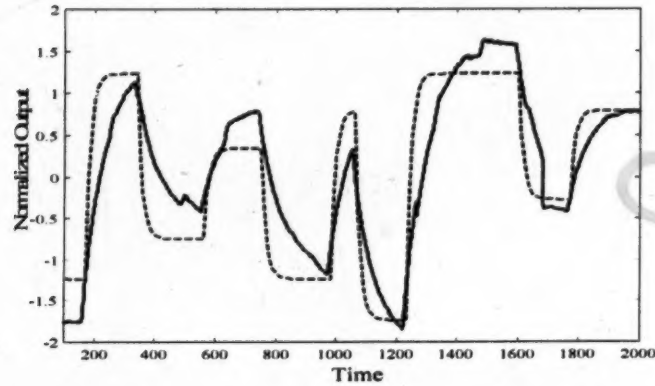


Figure 4.8: Comparison of simulation and real experimental data. Solid line presents the actual data and the dash lines show the data obtained from simulation.

process model of the CSTR reactor assuming the mechanistic equation is unavailable. The theory and algorithm of GA approach is discussed in detail in appendix B. The GA software applied in this study is provided by Zi-Jiang Yang at the Kyushu Institute of technology, Japan (2001).

Applying the same input-output data set obtained from experiment II and initializing the GA software (see appendixB) the process model is identified as follows:

Identified model using GAs:

$$y(t) = 0.998109 * y(t-1) + 0.00548206 * u(t-1)u(t-1) - 0.00352523 \quad (4.12)$$

The estimation result of conductivity applied on the cross validation data set is shown in figure 4.9

One of the distinct advantages of the proposed method is that without making *a priori*

Variable	Symbol	Nominal Value
Tank volume	V	1.795 dm ³
Reactor temperature	T	22c°
NaoH conc. in feed vessel	a_{μ}	0.1 mol/dm ³
ethylacetate conc. in feed vessel	b_{μ}	0.1 mol/dm ³
Volume feed rate of ethylacetate	F_b	60 dm ³ /s
operating point	$F_{in} = F_{out}$	60 dm ³ /s
Sampling period	Δt	100 s
Agitation speed	-	175 rpm

Table 4.5: Nominal conditions for process variables.

Input Signal	8 level uniformly distributed signal
Amplitude of excitation signal	60-80 Ml/min
Sampling time	100
Best formula (GMDH)	$a_1(t) = 0.013328 * F_a(t-1) + 1.0136 * a(t-1) - .00172 * a(t-1) * F_a(t-1) + 0.0005 * a^2(t-1) - .00238$
Linear Regression (LR)	$a_1(t) = 0.0924 * F_a(t-1) + 0.8962 * a(t-1) - 0.1253 * a(t-1) * F_a(t-1) + 0.1628 * a^2(t-1)$
Standard deviation of the identified coeff. by LR	$F_a(t-1): 0.0162$ $a(t-1): 0.0186$ $F_a(t-1) * a(t-1): 0.0259$ $a^2(t-1): 0.0238$

Table 4.6: Summary of identification results for the CSTR reactor, using an eight level uniformly distributed signal.

assumptions about the actual model, the structure and complexity of the model is evolved automatically.

Comparing equation 4.11 and 4.12 on the other hand, reveals that LR based model is a better understood model which can be explained by physical law of the process. Similar to GMDH technique, Genetic Algorithm provides an empirical model in which the estimated conductivity at the present time is highly correlated to the value of previous instant and the coefficient corresponding to the lagged output variable has significant effect in predicting the conductivity in the reactor.

The lack of sensitivity of the GA-based model to changes in various inputs shows that there are some limitations in the provided code. This follows the investigation of Zi-Jiang Yang (2001) that when dealing with simple non-linear systems, there is no significant advantage of the provided GA code compared to the conventional methods. It should be



Figure 4.7: Prediction results obtained from experimental data. The solid line is the normalized actual output and the dots represent the estimated values from equation 4.9.

emphasized that the provided GA code is not a commercial software and is just for basic investigation.

4.7 Conclusion

In this chapter, the issues of experimental design to the development of nonlinear inferential models of chemical systems are discussed. As an example of a nonlinear process, a pilot scale continuous stirred tank reactor, has been used to highlight the important identification issues related to the design of input sequence.

Designing an appropriate multi-level sequence and combining the knowledge of the CSTR process with Linear Regression (LR) algorithm, we are able to present a non-linear semi empirical dynamic model close enough to the mechanistic equation. It is shown that applying a pure "black box" technique such as Group Method of Data Handling provides a model, which might not be sufficient for describing the process information. However, this technique might be useful enough for prediction purpose when due to the poor understanding of the complex physicochemical processes, detailed first principles based model is unavailable.

Reviewing the application of Genetic Algorithm, Symbolic Regression has been applied as an alternative search technique in identifying the model structure when little information of the process is available. Final results revealed that LR based model gives a better estimation of the key quality variable compared to the GA and GMDH models.

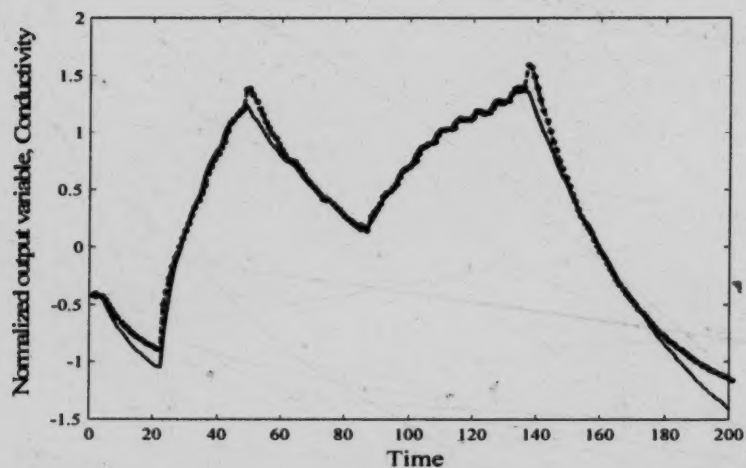


Figure 4.8: Prediction results obtained from experimental data. The solid line is the normalized actual output and the dash lines represent the estimated values from equation 4.9

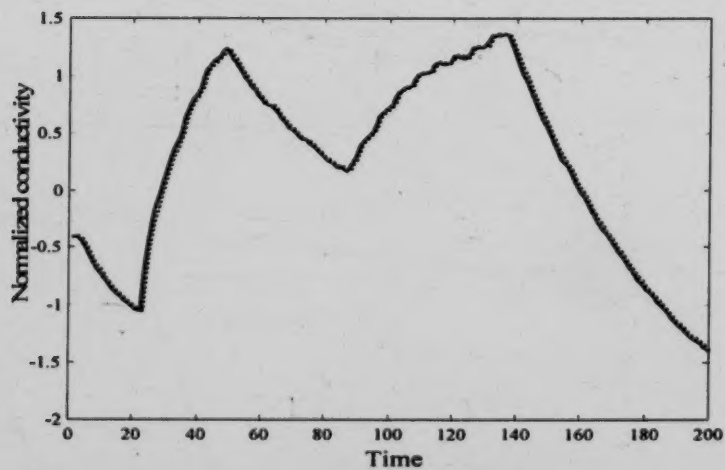


Figure 4.9: Prediction results obtained from the non-linear model provided by the Genetic Algorithm. The solid lines are the output from the CSTR and the dash lines represent the model prediction.

Chapter 5

Data Driven Model-on-Demand Approach

5.1 Abstract

In this chapter, the concept of data driven approach, model-on-demand (MOD), is introduced as another alternative solution to identify the nonlinear system, CSTR reactor. This technique provides a nonparametric estimation of nonlinear regression functions on the basis of availability of large datasets as the need arises. Reviewing the theory behind the MOD approach, we demonstrate the application of this identification tool on the CSTR reactor. The idea of MOD is expanded to the multi rate system to obtain the estimation of the output variable assuming that the input output data have different sampling rates.

5.2 Introduction

As discussed in previous chapters, one important issue in non-linear inferential modeling is identifying the model structure of the process. It was indicated that the global techniques such as nonlinear regression and neural network compress all available information into a compact model. Therefore, these methods become less attractive to deal with, when number of data points increases and the associated optimization problem becomes more complex.

The data driven approach, Model-on-Demand (MOD), is an alternative tool in system identification, which provides a nonparametric estimation of nonlinear functions on basis of available "large" datasets. In MOD estimation all observations are stored on a database and the models are built "on demand" as the actual need arises[?]. This technique uses only relevant and a small portion of data in the neighborhood of the new input to predict the key variable and determine a model as needed.

Adapting the number of data and assigning the relative weighting to the data points of interest, the variance/bias tradeoff is optimized locally and therefore the performance of model estimation will be improved. Another advantage of MOD cited in[?] is that unlike global modeling, there is no problem with adding new observation to the data set and this is due to the reason that there will be no estimation until a query time. Data sorting complexity and availability of large data sets as a necessary requirement, however, are the main drawbacks of this tool.

In this study, the theory behind model on demand is first studied. The application of this approach on the CSTR reactor is presented next. After comparing the prediction results using different techniques and different scenarios, we expand the idea of MOD to a simple multi rate system to obtain an estimation of the output variable assuming that the input-output data have different sampling rates. Finally, we discuss the results of the case study, taking into account the effort required to develop global models versus the MOD predictor.

5.3 Theory

The problem of identifying a nonlinear dynamical system described by

$$\begin{aligned}z(t) &= f(\phi(t)) \\ y(t) &= z(t) + e(t)\end{aligned}$$

where $e(t)$ and $\phi(t)$ are measurement noise and the regression vector respectively, has been solved traditionally using the global modeling methods such as nonlinear regression analysis. Another standard and traditional approach referred as *gain scheduling* [3] is to linearize the system at a discrete number of operational points parameterized by the scheduling variables and compute the linear model or controller for each of these operational points. The method we are addressing in this work is suggested by A. Stenman (1999) [50]. The idea is to store all past data in a database and compute a new estimate $\hat{z}$ for each new $\phi(t)$. To achieve

this goal, the relevant data belonging to a small neighborhood around the defined operating point x is retrieved from the dataset and a weighted linear regression is performed on that subset with the assumption that $f(\phi(t))$ is linear locally around $\phi(t)$.

The question arises at this point is what to be defined as relevant data, nearest neighbor estimator and weight selection. Following sections discuss how these key parameters are defined in the MOD approach.

5.3.1 Selecting the Weights and Distance Function

In weighted local modeling, the data subsets close to the target value are usually referred to as neighboring classes. Depending on the distance between each subset and the target value, each class has different contribution in key variable estimation. One important feature that distinguishes the local modeling from its global counterpart is that incorporating a weighting in the criterion gives less value to the irrelevant points, which are far from the objective class. Among different techniques in choosing the appropriate weighting factors, the local polynomial technique has been employed¹.

It is assumed that the weights are implicitly specified by a multivariable kernel function defined as follows:

$$w_i(x) = K_H(X_i - x) = \frac{1}{|H|} K(H^{-1}(X_i - x)) \quad (5.1)$$

where H is the bandwidth. Reformulating the above equation, the weighting scheme can be decomposed into two separate mappings; one that maps the local data to distances,

$$d(X_i, x) = \|X_i - x\|_M \quad (5.2)$$

where, $\|\cdot\|_M$ denotes a scaled vector norm, and another that maps the scaled distances to weights,

$$w_i(x) = K\left(\frac{d(X_i, x)}{h}\right) \quad (5.3)$$

Where, x is the target value and $K(\cdot)$ is a kernel function with following properties:

$$\int K(u)du = 1, \int uK(u)du = 0$$

Depending on the practical and theoretical considerations, a wide range of kernel functions is applicable. The common kernels used in this work are Gaussian kernel $k(u) = (\sqrt{2\pi})^{-1} \exp(-u^2/2)$ and tri cube kernel $K(u) = \frac{70}{80}(1 - (|u|)^3)^3_+$, where $(\cdot)_+$ denotes the positive part. It turns out that the local polynomial regression depends on the distance function $d(X_i, x)$ through the weighting $w_i(x)$ and therefore, the choice of distance function is an important design issue in local model estimation.

In general there are different types of distance functions depending on how much the regressors are stretched or shrunk. As shown in equation 5.2, the choice of the element

¹For other possible approach in choosing an appropriate smoothing weights see [50]

M will change the shape and the orientation of the neighborhood, which might affect the accuracy of the prediction. Among commonly used choices in the context of local modeling [4], we use the Euclidean distance function with M as identity matrix, defined as follows:

$$d(X_i, x) = \|X_i - x\| = \sqrt{(X_i - x)'(X_i - x)} \quad (5.4)$$

5.3.2 Bandwidth

Another important issue in MOD approach is bandwidth estimation. Choosing a small bandwidth makes the polynomial fitting mainly dependent on the data points close to the target value x and therefore produce a noisy estimation. A large bandwidth, on the other hand, results an estimate which is very close to global linear fitting.

In practical application, the available data set is an important source in estimating an appropriate bandwidth. This is usually referred to as *(data-driven)* bandwidth.

The two broad methods for selecting the bandwidth are introduced in [35] and [50] as follows:

Classical Methods: Using the classical tools applied in parametric modeling for minimizing the mean square error "MSE". Cross validation, AIC, FPE and Mallows C_p are such typical examples.

Plug-in Methods: Estimating the unknown quantities from the data and "plugging in" to compute an asymptotically optimal bandwidth [50].

In this study, we choose the first approach, classical technique, as a way of getting the acceptable bandwidth.

5.3.3 Classical Methods, Variance/Bias Balance

The classical method is based on the quality of the estimator as a function of the bandwidth parameter and optimizing this measure with respect to the bandwidth. Therefore, to assess the performance of the local fit, a criterion is required. Depending on the application, different criteria can be employed. For the sake of simplicity, we use the localized final prediction error criterion defined as follows:

$$FPE(x, K) = \frac{\sum_{i \in \Omega_k(x)} \omega_i(x) (Y_i - \hat{m}(X_i, k))^2}{tr(W_k)} \times \frac{2tr(W_k) + \alpha tr((X_k^T W_k X_k)^{-1} (X_k^T W_k^2 X_k))}{2tr(W_k) - \alpha tr((X_k^T W_k X_k)^{-1} (X_k^T W_k^2 X_k))} \quad (5.5)$$

where, the first part of the above equation is equivalent to the mean square error and the variance penalty ($\alpha \geq 2$) is an arbitrary penalty applied on the variance term of the criterion when the small neighborhood is chosen. This reduces the possibility of finding false feature in local estimation. Parameters X and W denote the design and weight matrix respectively.

The presented localized goodness-of-fit estimates the quality of the fit for a given neighborhood size k . This estimation is applied on different k and depending on the lowest cost

for a given goodness-of-fit measure, which provides a balance between bias and variance error, the optimal k is chosen automatically. For other possible criteria the reader is referred to [50].

5.3.4 Model-on-demand Algorithm

The general algorithm of the MOD estimator based on local regression is summarized in this section. Providing the historical observations $(Y_i, X_i)_{i=1}^N$, an estimation point x , type of kernel and distance function and goodness-of-fit measure, the predicted value $\hat{m}(x)$ and its variance will be estimated by following the procedures indicated below:

- Sorting the available data in ascending order according to the distance from x
- Performing weighted regression fit at a very small bandwidth h_0 , close to the smallest bandwidth for which a well defined estimate is obtained.
- Increasing the bandwidth exponentially according to

$$h_i = c_h \cdot h_{i-1}$$

where $C_h > 1$. As suggested by Loader (1997)[36], this coefficient can be calculated as follows:

$$C_h = 1 + \frac{0.3}{d}$$

This procedure is repeated until a goodness-of-fit cost fails at a low significance level or a maximum neighborhood size k_{max} is exceeded [50].

- Selecting the K_{opt} based on the lowest cost of goodness-of-fit.
- Choosing the corresponding recorded parameter vector $\hat{\beta}$ at k_{opt} to estimate the target object $\hat{m}(x)$

5.4 Illustrative Examples of MOD Algorithm

The proposed modeling strategy has been successfully applied to the pilot scale CSTR reactor presented in the last chapter. To prove the feasibility of this approach, we apply different scenarios covered in the following sections.

5.4.1 Identifying the simulated CSTR system using a linear model

Using the 8-level excitation signal and simulating the system, a training set of data for 1000 observations has been produced. Neglecting the nonlinear term in CSTR system, a global linear model was constructed based on linear regression technique. The prediction result is tested on another 500 simulated data points. Using the same training and testing data sets, a robust and reliable prediction was obtained using the MOD algorithm. The estimations of NaOH conductivity using global regression and local fit is compared in figure 5.1. As it

is shown, the local fit estimation is very close to the actual data. Note that the local fit is based on a linear model.

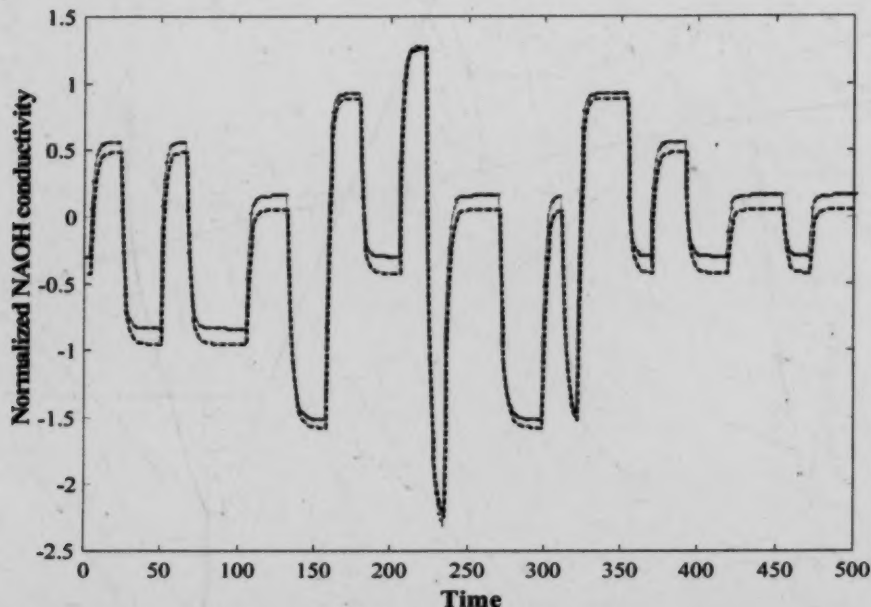


Figure 5.1: Prediction results of NAOH conductivity obtained from global linear regression and MOD tool . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.

5.4.2 Evaluating MOD robustness using contrived nonlinear model structure

To examine the robustness and reliability of the MOD technique, it is assumed that the exact model structure is unknown and the model consists of the linear and cross term of lagged input output variables. Employing the global regression and weighted local fit on the same set of data, a poor estimation is obtained by global modeling while the estimated product conductivity based on MOD fit is almost identical to the actual data. The prediction result of both techniques is shown in figure 5.2.

5.4.3 MOD prediction performance in the presence of disturbances

Applying the band limited white noise on the previous set of data, we evaluate the performance of the model on demand approach by comparing the global regression and local fit following the last two scenarios. Figure 5.3 and 5.4 show the estimation results of the noisy data using linear and contrived nonlinear model structure respectively.

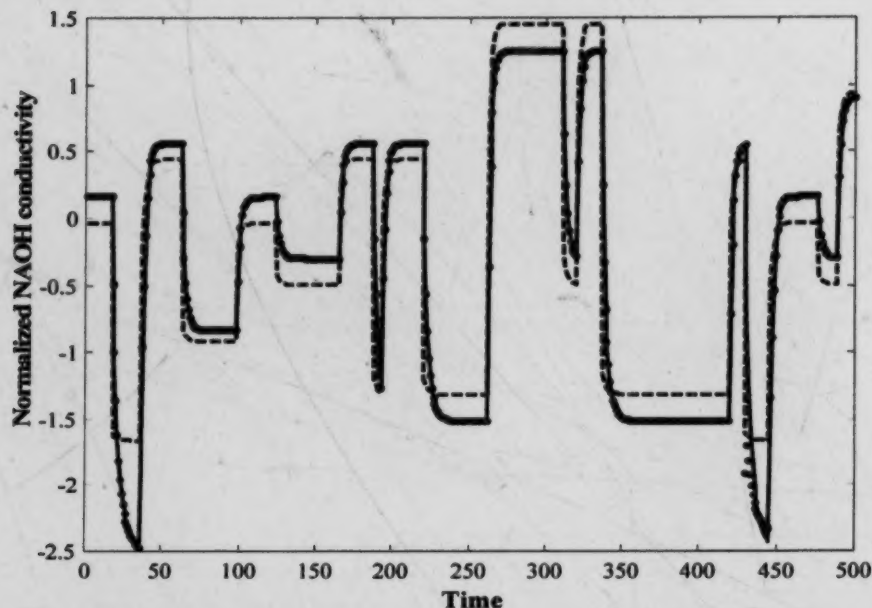


Figure 5.2: Prediction results of NAOH conductivity obtained from global regression and MOD tool using a false model structure . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.

It is shown that the MOD estimator does provide a better prediction compared to global regression.

5.4.4 Comparison of prediction performance of CSTR system using ANNs, GMDH and MOD technique

Employing the nonlinear identification tool ANNs, on the same training and testing data set including band limited white noise, another estimation for the NAOH conductivity is obtained.

To compare the performance of the candidate nonlinear black box techniques, GMDH, MOD and ANNs, all prediction results are plotted in figure 5.5. As presented in this figure, the GMDH and MOD provide closer estimation to the actual data compared to the one obtained from ANNs.

5.4.5 Evaluation of MOD technique using experimental data

Performing the local fit on the *experimental* data set presented earlier, the effectiveness of proposed identification algorithm is evaluated. The data obtained from conducted experi-

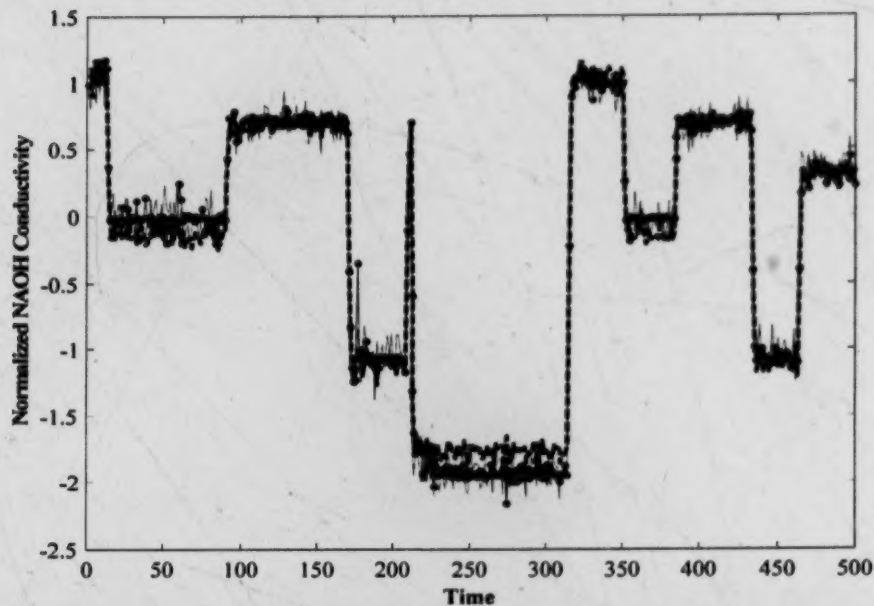


Figure 5.3: Prediction results of NaOH conductivity from global linear regression and MOD tool applied on the noisy data set . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.

ments are shown in figures 4.7 and 4.8. The first set of data are used for model identification while the second set treated as testing data set.

As shown in figure 5.6 the predicted response using the MOD algorithm is able to fit the actual set of data well. The presented simulation and experimental results show the distinct feature of model-on-demand approach and its ability to predict the process output very effectively while being robust to the disturbances.

5.5 Multi-rate system

This section extends the application of model-on-demand methodology to the systems in which the quality control and measured variable are sampled at a different rate.

We consider the pilot scale CSTR reactor as a "linear" multi rate system with the input sampling rate of every time unit and output sampling rate of every two units. The general parametric model for one, two and three step ahead prediction of the system can be presented in equations 5.6, 5.7 and 5.8 as follows:

$$y(t+1) = \alpha y(t) + \beta u(t) \quad (5.6)$$

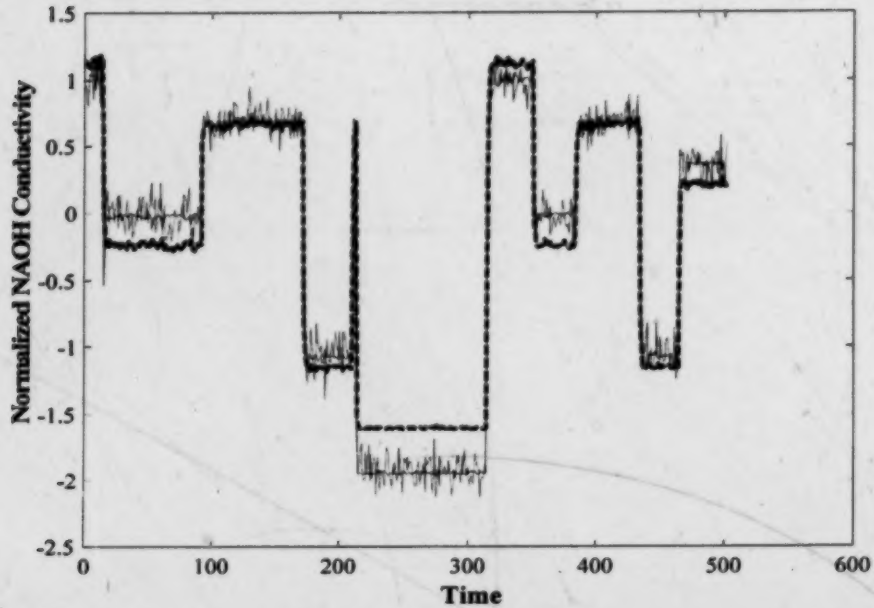


Figure 5.4: Prediction results of NAOH conductivity from global regression and MOD tool using a false model structure applied on the noisy set of data. The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.

$$y(t+2) = \alpha y(t+1) + \beta u(t+1) \quad (5.7)$$

$$y(t+3) = \alpha y(t+2) + \beta u(t+2) \quad (5.8)$$

The information of $y(t+1)$ and $y(t+3)$ is available while the observations for the $y(t+2)$ is unknown. In this study, the ratio between output/input data is 2 ($M = 2$). The objective is to estimate the unknown parameters α and β and build a model for output variable based on each unit sampling time. To achieve this goal, we substitute 5.7 in 5.8 and obtain the following equation in which all available data is used.

$$y(t+3) = \alpha^2 y(t+1) + \alpha \beta u(t+1) + \beta u(t+2) \quad (5.9)$$

Equation 5.9 can be solved using the ordinary regression technique. Having two unknown variables and three available equations, the parameters α and β are optimized using the

Black box technique	Mean Square Error
GMDH	5.6005
Neural Networks	8.4960
Global linear regression	12.4608
Model-On-Demand	6.01

Table 5.1: Comparison of mean square error (MSE) of different algorithms applied to data set with band limited white noise

non-linear least square estimation. The result can be easily extended for an arbitrary M . However, the complexity is increased quickly with the increase of output/input sampling ratio. An alternative approach to the modeling of the multi rate system is to transfer the SISO system into MIMO system using the lifting technique as discussed in Chen and Francis.

The technique introduced in this work is identical for both global and local modeling except that in the local estimation this procedure is repeated for every single data point and therefore the estimated parameters α and β are different for each observation.

Subsequent analysis of CSTR simulations will demonstrate that applying the presented multi rate sampling strategy along with the model on demand approach for parameter estimation, significant improvements in prediction performance can be engendered.

5.6 Simulated examples using multi rate operations

Given the set of data presented in first subsection of 5.4.1, we first follow equation 5.9 and perform global regression to estimate the parameters α , $\alpha\beta$ and β . Optimizing the parameters α and β , the two step ahead prediction components can be computed using equation 5.7. These procedures provide a prediction based on global linear regression.

Incorporating the feature of MOD local estimation into the multi rate strategy and repeating the same procedures, a better estimation for each new observation can be obtained. The result of both local and global estimation combined with multi rate algorithm is shown in figure 5.7.

Performing the same methodology to the data points including band limited white noise,

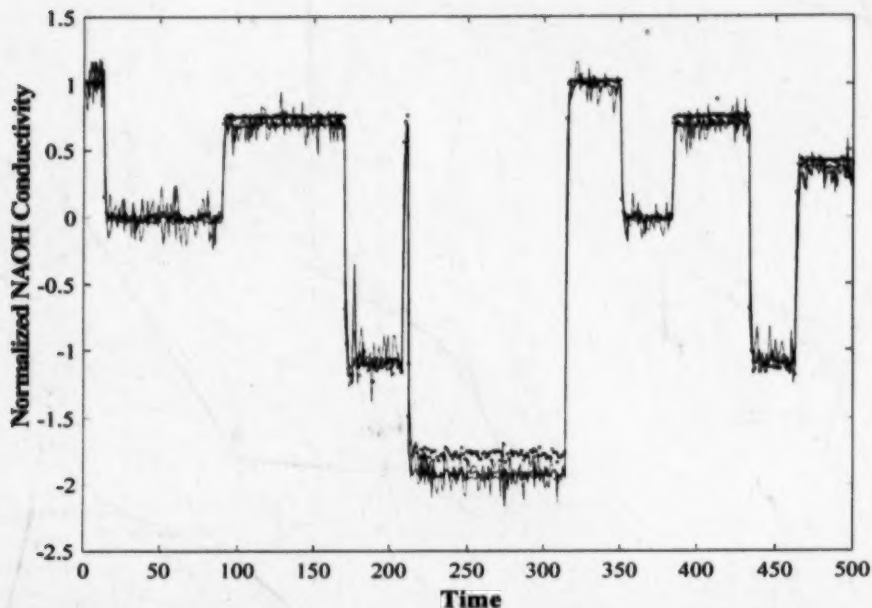


Figure 5.5: Prediction results of NAOH conductivity from GMDH, MOD and ANNs tool applied on the noisy set of data. The solid lines are the simulated output from the CSTR; the GMDH and MOD are shown with dashes and the dots respectively. The square sign represents ANNs.

we obtain the prediction results which are shown in figure 5.8.

The results from these simulations reveal that the model on demand approach is the promising candidate in predicting the multi rate systems.

5.7 Conclusion

We have studied the problem of system modeling using data stored in a database. A model-on-demand approach is used to estimate the process model of the CSTR reactor. The MOD predictor is formed locally around the current working point such that the pointwise error is minimized and therefore, a good bias/variance tradeoff is achieved.

The application of MOD algorithm is expanded to the multi rate systems in which the input data is sampled faster than the output signal. The simulation results illustrated the effectiveness and feasibility of the algorithm.

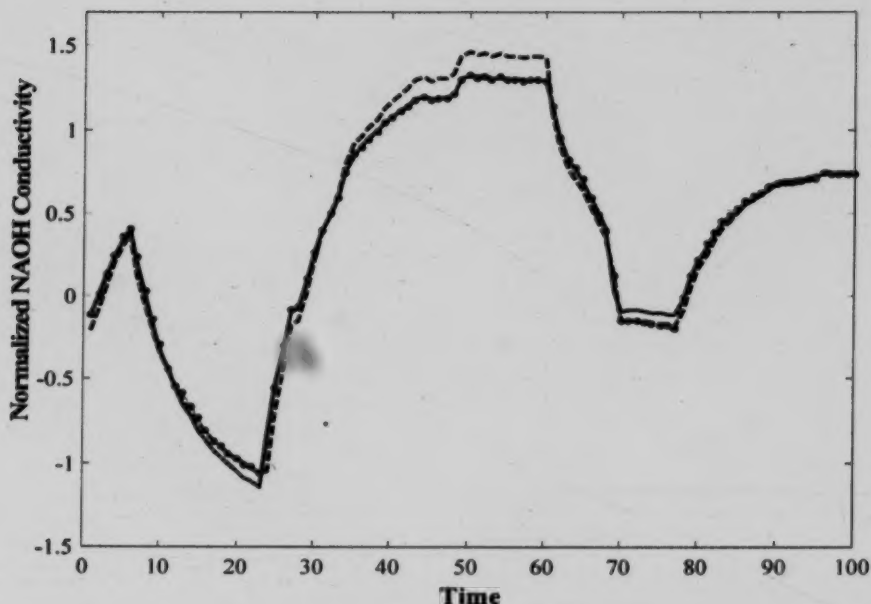


Figure 5.6: Prediction results of NAOH conductivity from MOD tool applied on the experimental set of data. The solid lines are the output from the pilot scale CSTR reactor. The solid lines are the actual output from the CSTR and the dots represent the MOD based prediction.

Model structure	MSE Global	MSE local
Linear model without disturbance	1.3087	0.2769
False nonlinear model without disturbance	44.6171	4.2009
Linear model on the "Experimental" data set	0.6180	0.1160
Linear model with band limited white noise	12.4608	7.6660
False nonlinear model with band limited white noise	23.3917	6.8762
Linear model without disturbance on the Multi rate system	0.9196	0.2831
Linear model with band limited white noise on the Multi rate system	9.3816	8.3564

Table 5.2: Comparison of mean square error (MSE) of the discussed scenarios

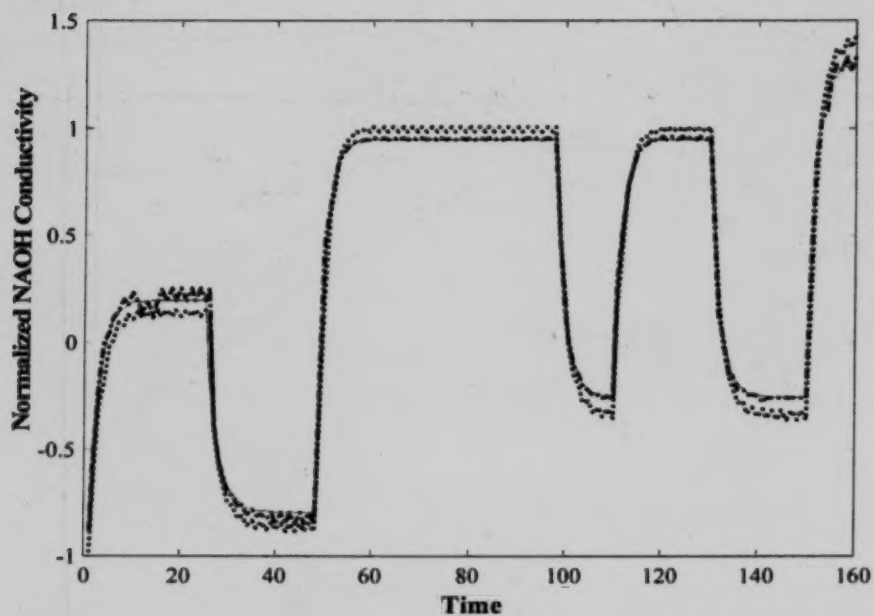


Figure 5.7: Prediction results of NAOH conductivity obtained from global linear regression and MOD tool applied on the multi rate system . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.

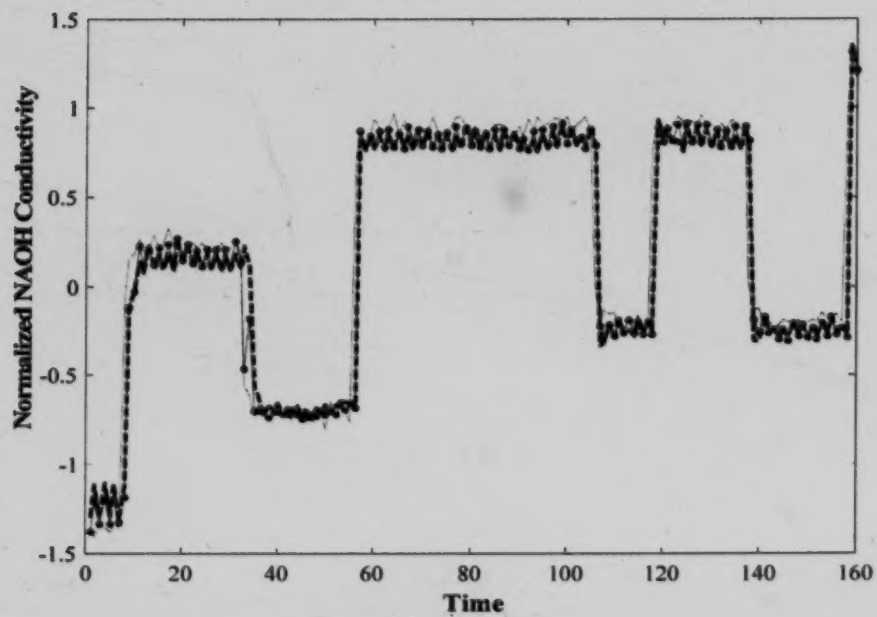


Figure 5.8: Prediction results of NAOH conductivity from global linear regression and MOD tool applied on the noisy data set. The data is obtained from the simulated multi rate system . The solid lines are the simulated output from the CSTR; the global regression is shown with dashes and the dots represent the MOD based prediction.

Chapter 6

Conclusions

This thesis has explored the applications of first principles laws, multivariate statistical and mathematical tools as well as issues in experimental design for identification of industrial and pilot-scale processes.

The industrial case study, a Fluid Coker, has been considered in Chapters 1 and 2 in order to visualize the use of presented tools for real application. The semi-empirical based model obtained in Chapter 1 has yielded a reasonably good prediction of top distillation point of the Coker at Syncrude Canada Ltd. To enhance the prediction performance, different identification algorithms including partial least square (PLS) regression, group method of data handling (GMDH) and artificial neural networks (ANNs) have been presented in Chapter 2. The identified nonlinear components from Chapter 1 were used as compensating elements in the design of the PLS based modeling. It is shown that the feasibility of the PLS approach depends on the availability of the model structure under consideration. To identify the *optimal* model under the condition that *a priori* information of the process is not easily at hand, GMDH tool has been applied on the same input/output data. The results have indicated that the GMDH tool provides a better estimation while the model structure is self-organized. Neural networks have also provided reasonably successful prediction, however, no explicit model is available by this tool. In Chapter 3, the issues of experimental design in the development of nonlinear inferential models of chemical systems were discussed. By designing an appropriate multi-level sequence and performing linear regression (LR), GMDH and genetic algorithms (GAs), a final LR based model has been provided for the nonlinear pilot scale CSTR reactor. The model on demand (MOD) approach was described in Chapter 4. This tool has been applied on the pilot scale CSTR reactor presented in Chapter 3. It has been shown here that the MOD method is much more reliable compared to other global methods. The MOD technique was also extended to the domain of the simple multi rate system.

6.1 Contribution of this thesis

The key contributions of this study can be listed as follows:

- One industrial case study involving first principles and black box modeling is presented to demonstrate the utility of these techniques for inferential model development.
- Superiority of the first principles modeling (FPM) over pure black box techniques was established via extensive comparison and study.
- Constructing a hybrid PLS/first principles model, it has been shown that such a model can outperform the conventional one.
- An identification tool, group method of data handling, has been proposed when the mechanistic equation is unavailable.
- A uniformly distributed M-level signal has been proposed to the identification of a nonlinear system. Theoretical developments were supported by simulation and experimental studies.
- Some practical extensions were made to the model on demand strategy. A new algorithm was suggested in conjunction with the MOD approach to handle the problem of system modeling when the input data are sampled faster than the output signal.

6.2 Recommendation for future work

- As mentioned earlier, there have been considerable issues in the identification of nonlinear systems in the literature, which have not been completed and require further study. As an example, additional improvements in nonlinear inferential modeling can be expected if better ways of design of experiments are incorporated.
- Investigation of the performance of the MOD approach on the model predictive controller may be a potential research direction.
- It may also be interesting to apply MOD to a more general multi rate system using the lifting technique.

Bibliography

- [1] American Petroleum Institute (1996). Technical data book-petroleum refining
- [2] Aström K.L. and McAvoy T.J. (1992). Intelligent control. *Journal of Process Control*, NO. 2, 115-126
- [3] Aström K.J. and Wittenmark B. (1989). Adaptive Control. Addison Wesley
- [4] Atkeson C.G., Moore A.W. and Schall S. (1997). Locally weighted learning for control. *Artificial Intelligence Review* 11 (1-5), 75-113
- [5] Barker H.A. and Davy R.W. (1978). Second order Volterra Kernel measurement using pseudorandom ternary signals and discrete fourier transformations. *Int. J. of Control* 27, 277-291
- [6] Barker H.A., Obidegwu S.N. and Pradisthayon T. (1972). Performance of antisymmetric pseudorandom signals in the measurement of Volterra kernels by cross-correlation. *proceedings IEE* 119, 353-362
- [7] Bell, H.S. (1989). American Petroleum Refining
- [8] Braun M.W., Rivera D.E., Stenman A. and Foslien W. (2000). Comparison of global nonlinear models and model on demand estimation applied to identification of ARTP wafer reactor. *Conference paper* Presented at SYSID 2000, Santa Barbara, CA
- [9] Chambers L. (1995), Practical Handbook of Genetic Algorithms Applications, Vol. 1
- [10] Chen S., Billings S.A. and Luo W. (1996). Orthogonal least squares methods and their application to non-linear system identification *INT. J. CONTROL*. Vol. 50, No. 5, 1873-1896
- [11] Farlow S.J. (1984). Self-Organizing Methods In Modeling GMDH Type Algorithms
- [12] Fausett L. (1994). Fundamentals of Networks, ARCHITECTURES, ALGORITHMS AND APPLICATIONS.
- [13] Felder R.M. and Rousseau R.W. (1989). Elementary Principles of Chemical Processes

- [14] First IEEE International Conference on Genetic Algorithms (1995). International Conference on Genetic Algorithms in Engineering Systems: Innovations and applications
- [15] Fontes C. and Mendes M.J.(2000). Modeling of an Ethylene Polymerization Slurry Reactor, Neural versus Phenomenological Models, *IFAC* June 14-16 623-628
- [16] Furuhashi T. and Uchikawa Y. (1995). Lecture notes in Artificial Intelligence, Fuzzy Logic, Neural Networks and Evolutionary Computation
- [17] Geladi, p., and Kowalski B.R. (1986). Partial Least Squares Regression: A Tutorial *Analytica Chimica Acta*, 185, 1-17
- [18] Giron-Sierra J.M., Anders-Toro B.D, Blanco P.F. (2000), Application of Genetic Algorithms to determine parameters of process models. *ADCHEM* Pisa, Italy
- [19] Godfrey K. (1993). Introduction to perturbation signals for time-domain system identification. pp. 1-59
- [20] Godfrey K. (1969). The theory of the correlation method of dynamic analysis and its application to industrial processes and nuclear power plant. *Meas. and Control* 2
- [21] Goldberg D.E. (1989), Genetic Algorithms, International Student Edition
- [22] Gron-Sierra J.M., Andres-Toro B.de, Fernandez Blanco P. (2000), Application of Genetic Algorithms to Determine Parametes of Process Models. *IFAC, ADCHEM* Pisa, Italy
- [23] Haber R. and Unbehauen H. (1990). Structure Identification of Nonlinear dynamic systems- A survey on Input/Output approaches *Automatica* Vol. 26, No. 4, 651-677
- [24] Hoeskuldsson A. (1988). PLS Regression Methods *Journal of Chemometrics* Vol.2, 211-228
- [25] Ikeda S., Ochiai M. and Sawaragi Y. (1976). Sequential GMDH algorithm and its application to river flow prediction *IEE Transactions on Systems, Man and Cybernetics* Vol. Smc-6, No. 7
- [26] Ivakhnenko A.G., Wunsch D. (1990). Inductive Sorting-out GMDH Algorithms with polynomial Complexity for Active Neurons of Neural Network
- [27] Juditsky A., Hjalmarsson H., Benveniste A., Delyoun B., Ljung L., Sjoberg J. and Zhang Q. (1995) Nonlinear Black-box Models in System Identification, Mathematical Foundations *Automatica* Vol.31, No.12, 1725-1750
- [28] Kaspar M.H. and Ray W.H. (1992). Chemometric Methods for Process Monitoring and High-Performance Controller Design *AIChE Journal* Vol.38, No.10 1593-1608

- [29] Katz D.L. (1959). Handbook of natural gas engineering
- [30] Kortmann M., Janiszowski K. and Unbehauen H. (1988). Application and comparison of different identification schemes under industrial conditions *INT. J. CONTROL*. Vol. 48, No. 6, 2275-2296
- [31] Lakshminarayanan S. (1997). Process Characterization and Control using Multivariate Statistical Techniques *Ph.D. thesis*, University of Alberta
- [32] Leffler W.L. (1979). Petroleum Refining
- [33] Lisboa P.J.G. and Taylor M.J. (1993). Proceedings of the Workshop on Neural Network Applications and Tools
- [34] Ljung L. (1987). System Identification: Theory for the User
- [35] Loader C.R. (1995). Old faithful erupts: Bandwidth Selection Reviewed. *Technical report. AT&T Bell Laboratories*
- [36] Loader C.R. (1997). Local fit: An Introduction. *Technical report. AT&T Bell Laboratories*
- [37] MAO K.Z. and Billings S.A. (1997). Algorithms for minimal model structure detection in nonlinear dynamic system identification *INT. J. CONTROL*. Vol. 68. No. 2, 311-330
- [38] Maxwell J. B. (1968). Data book on hydrocarbons, Application to Process Engineering
- [39] McKay B., Willis M.J. and Barton G.W. (1995). On the Application of Genetic Programming to Chemical Process Systems *IEE*.
- [40] McKay B., Willis M., Montague G. and Barton G. (1995). Using Genetic Programming to Develop Inferential Estimation Algorithms. *European workshop proceedings*
- [41] Mikhail E.M. and Gracie G. (1996). Analysis and Adjustment of Survey Measurements
- [42] Nelson W. L. (1958). Petroleum Refinery Engineering
- [43] Poli R., Nordin P., Langdon W.B. and Fogarty T.C. (1999), Proceedings lecture Notes in computer science, Genetic Programming. *Second European workshop, EuroGP'99*
- [44] Pottmann M., Unbehauen H. and Seborg D.E. (1993). Application of a general multi-model approach for identification of highly nonlinear processes- a case study. *INT. J. Control* Vol. 57, No. 1, 97-120
- [45] Sattler K. and Feindt H.J. (1995). Thermal Separation Processes, Principle and Designs
- [46] Singh B., Naps T.L. (1985), Introduction to Data Structures

- [47] Skrzypek J. (1994). Neural Network Simulation Environments
- [48] Söderström T. and Stoica P. (1989). System Identification. Prentice-Hall, Hemel Hempstead, Hertfordshire, UK
- [49] Stenman A., Gustafsson F. and Ljung L. (1998). Just In Time Models for Dynamical Systems
- [50] Stenman A. (1999). Model on Demand: Algorithm, Analysis and Applications *Ph.D. Thesis*, Linköping University, Sweden
- [51] Syncrude Canada Ltd. (2000). Documentation of Upgrading unit
- [52] Tenenbaum A.M. and Augenstein M.J. (1981). Data structures using pascal
- [53] Wechsler H. (1992). Neural Networks for perception: Computation, Learning and Architectures. Vol. 2

Appendix A

Nomenclature List

F : Total volume feed rate (dm^3/s)

F_a : Volume feed rate of sodium hydroxide (dm^3/s)

F_b : Volume feed rate of ethyl acetate (dm^3/s)

V : Volume of reactor (dm^3)

a_0 : Sodium hydroxide conc. in mixed feeds (mol/dm^3)

a_1 : Sodium hydroxide conc. in reactor at time t (mol/dm^3)

k : Specific rate constant

Appendix B

Introduction to Genetic Algorithms

Being a part of evolutionary computing, genetic algorithm (GA) is a rapidly growing technique in a variety of engineering fields and different application areas like optimization and system modeling. GA is the study of artificial systems that has the characteristic of natural living systems. This algorithm is based on the fact that those least suited to the current environment die, while the best live on, to produce the next generation. In the same environment, the best of each generation increase its fitness over successive generations until an optimum solution is achieved.

Genetic algorithms operate on populations of strings¹, with the string coded to represent some underlying parameter sets. Reproduction, crossover and mutation are applied to successive string populations to create new string populations. These operators involve random number generation, string copying, and partial string exchanging.

To review the genetic algorithm and the function of its operators in more detail, we discuss one of the application of this technique, developing an inferential process model using symbolic regression. Like GMDH, symbolic regression determines the appropriate structure and complexity of the model as well as the best set of parameters. This approach is discussed in more detail in the following sections.

B.1 Performing Symbolic Regression, an Application of Genetic Algorithm

Regression is a techniques used quite frequently to interpret time series relationship. It consists of finding the coefficients of a *prefixed* function such that the resulting function best fits the data points. However, the problem arises when finding good coefficients is impossible and the function, which performs the best result, is unknown.

¹The strings of an artificial genetic system are analogous to chromosomes in the natural and biological systems. The chromosomes are composed to genes, which are identical to feature, character or detector in GAs. Like the position of a gene in the chromosome, its locus, features may be located on different positions of the string.

A tree structured genetic algorithm is one of the applied methodologies which, unlike conventional regression, determines the structure and hence the complexity of the model. The term "symbolic" stresses the fact that we are not interested in finding optimum parameters (numbers) but optimum functions (expressions, symbolic representations).

In order to use GA to solve symbolic regression problems, we follow the procedures indicated below:

- After initializing and generating the first population of solutions to the numeric prediction problem, the next step in genetic algorithm is coding parameters. In order not to restrict the domain of the search and apply this technique to the general problems, tree coding is applied to represent the algebraic expressions.

Assuming the output variable y should be predicted by three inputs x_1 , x_2 and x_3 the tree structure of the initialized population, $\frac{(x_1 - x_2)}{3}$, $(x_3 + x_2) * (x_1 - x_3)$ can be presented as figure B.1. For a detail discussion of tree structure see Tenenbaum and Augenstein (1981).

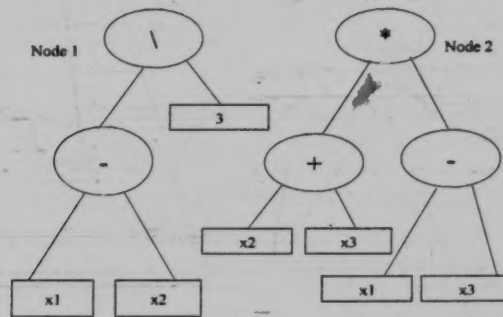


Figure B.1: Representation of numeric expressions using tree structures. (Parent Population)

- To evaluate the fitness of each individual tree and to measure the capability of each member in predicting the target value, we assign a fitness functions to each tree. One example of the fitness function introduced in the majority of literatures is the minimum value of the error function between the actual and predicted solutions.
- Initializing the parent population and assessing the fitness of each member, we apply the genetic operators, reproduction, crossover and mutation. From this stage, the evolution of the process can be started. Note that the choice of each operator is probabilistic.
- *Crossover* is randomly selection of a crossover point (link between nodes) in each parent and swapping the sub-trees laying below the crossover points. Selecting a crossover point is equivalent to selecting a random sub-expression.

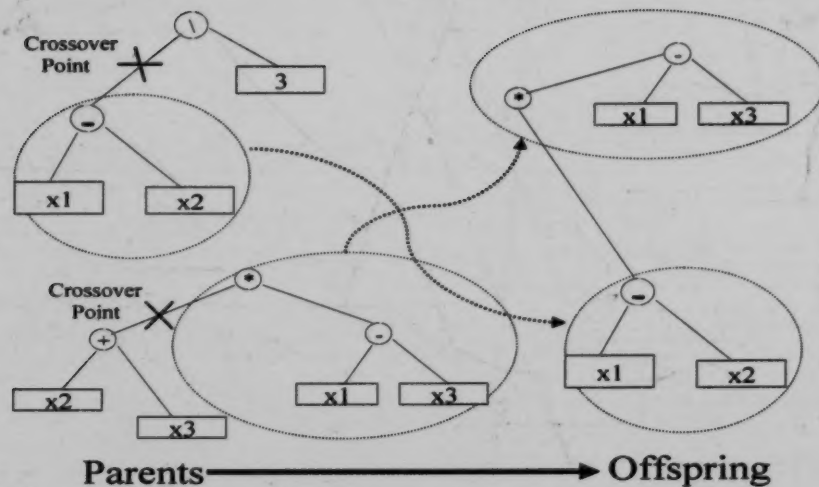


Figure B.2: Example of crossover operation.

- **Mutation** is randomly selection of a mutation point in a tree and substituting the sub-tree laying below such a point with a randomly generated sub-tree:

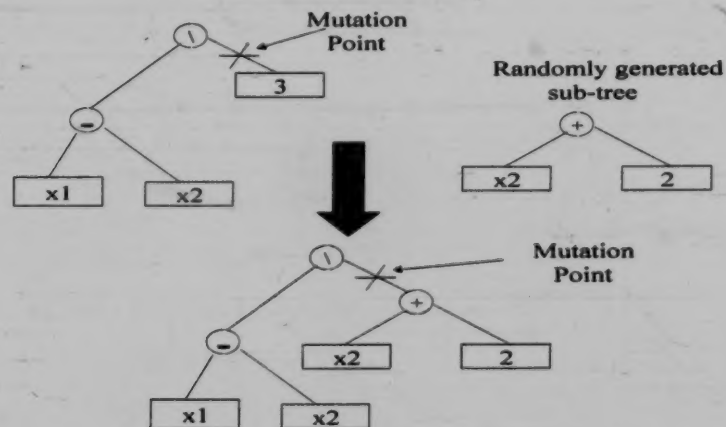


Figure B.3: Example of Mutation operation.

- Applying the genetic operators, we measure the fitness function of each new offspring to make the decision which members of the population should die. This cycle is then repeated until the maximum numbers of generations or other pre specified convergence criteria are reached.

These procedures identify the structure of the nonlinear model. Using any optimization method we are able to find the coefficients of the model at the same time and

reach the optimal solution. Reviewing the implementation aspects of tree structured symbolic regression, we might notice that there are several user specified parameters such as population size, the number of generation, crossover and mutation probability, available operators (+, -, *, ...) and their probability and finally the weighting variable of the fitness of each tree to prevent over fitting of the data and penalizing the formation of a large tree.

The general genetic algorithm can be summarized in figure B.4:

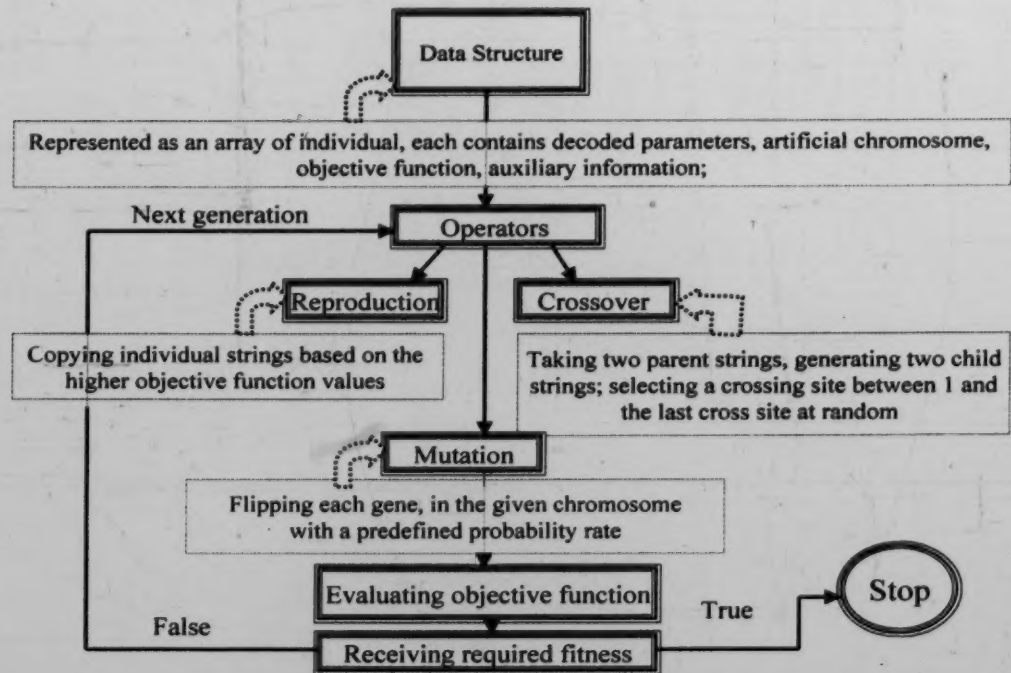


Figure B.4: Basic algorithm of the evolutionary computation, genetic algorithm in a flow chart scheme.